A cooperative crowdsourcing framework for knowledge extraction in digital humanities – cases on Tang poetry

Liang Hong, Wenjun Hou, Zonghui Wu and Huijie Han
School of Information Management, Wuhan University, Wuhan, China

Received 31 July 2019
Revised 15 November 2019
4 January 2020
5 January 2020
14 January 2020
Accepted 27 January 2020

Abstract

Purpose – The purpose of this paper is to propose a knowledge extraction framework to extract knowledge, including entities and relationships between them, from unstructured texts in digital humanities (DH).

Design/methodology/approach – The proposed cooperative crowdsourcing framework (CCF) uses both human–computer cooperation and crowdsourcing to achieve high-quality and scalable knowledge extraction. CCF integrates active learning with a novel category-based crowdsourcing mechanism to facilitate domain experts labeling and verifying extracted knowledge.

Findings – The case study shows that CCF can effectively and efficiently extract knowledge from multi-sourced heterogeneous data in the field of Tang poetry. Specifically, CCF achieves higher accuracy of knowledge extraction than the state-of-the-art methods, the contribution of feedbacks to the training model can be maximized by the active learning mechanism and the proposed category-based crowdsourcing mechanism can scale up the effective human–computer collaboration by considering the specialization of workers in different categories of tasks.

Research limitations/implications – This research proposes CCF to enable high-quality and scalable knowledge extraction in the field of Tang poetry. CCF can be generalized to other fields of DH by introducing domain knowledge and experts.

Practical implications – The extracted knowledge is machine-understandable and can support the research of Tang poetry and knowledge-driven intelligent applications in DH.

Originality/value – CCF is the first human-in-the-loop knowledge extraction framework that integrates active learning and crowdsourcing mechanisms; he human–computer cooperation method uses the feedback of domain experts through the active learning mechanism; the category-based crowdsourcing mechanism considers the matching of categories of DH data and especially of domain experts.

Keywords Crowdsourcing, Human–computer cooperation, Knowledge extraction, Digital humanities, Tang poetry

Paper type Research paper

1. Introduction

With the continuous development of information technology, researchers begin to use interdisciplinary research methods of digital humanities (DH) to open up a new paradigm for humanities research. Consequently, a great deal of DH data has been accumulated, such as subject databases, electronic archives, knowledge bases, webpages and so on. The multi-source heterogeneous data, which are difficult to read and understand by computers, increase the difficulty and workload of DH research. Therefore, it is necessary to extract machine-understandable knowledge from the data and organize the extracted knowledge into a knowledge graph to support DH research.

Tang poetry is the representative of traditional Chinese literature and one of the highest achievements of Chinese poetry creation. There are more than 50,000 poetries written by over 2,200 poets in the Tang dynasty, which have a far-reaching influence on Chinese culture and even world culture. At present, there are a large number of experts in China who conduct researches on Tang poetry and have made fruitful achievements (Li, 2010).

As one of the important fields of DH, Tang poetry has accumulated a large amount of data resources. However, these resources are scattered, sparse and lack effective organization.

Aslib Journal of Information Management
Vol. 72 No. 2, 2020
pp. 243-261

2050-3806
DOI 10.1108/AJIM-07-2019-0192

In the field of Tang poetry, knowledge extraction can provide a solution to transform multi-source heterogeneous data into "intelligent" linked data, i.e. entities and relationships between them. This, in turn, provides a solid foundation for knowledge association and reasoning, which supports DH studies and intelligent applications.

In this paper, we study the problem of extracting knowledge from unstructured texts of Tang poetry. However, it is not an easy task because of the large scale of data and the unique characteristics of Tang poetry. Firstly, Tang poetry is a type of ancient Chinese text that has unique terms of words, sentence patterns, grammar and rhyme schemes. Secondly, state-of-the-art knowledge extraction techniques such as machine learning and deep learning are lack of training instances and prior knowledge (Alani *et al.*, 2003), thus cannot be directly applied to such humanities research. Last but not least, knowledge extraction relying on domain experts is costly and not scalable to a large amount of data. Although some studies (Plaisant, 2006) combined the efforts of domain experts and computer systems in the DH field, they still cannot solve the problem of unmatchable speed and scale between computer processing (e.g. machine learning) and human work. For instance, machine learning algorithms can generate hundreds of thousands of training results in 1 min, while domain experts can only label several of them during such a short time. Moreover, because of the specialty and universality of Tang poetry, each domain expert is good at a small portion of whole domain knowledge.

To address the above challenges, we propose a cooperative crowdsourcing framework (CCF) to extract knowledge from Tang poetry data effectively and efficiently. CCF improves the quality of extracted knowledge by introducing feedbacks of domain experts through active learning mechanisms. Meanwhile, the quality and scalability of labeling by domain experts are also improved by introducing a category-based crowdsourcing mechanism.

CCF contains Input Engine, Machine Extraction Engine and Crowdsourcing Engine. In Input Engine, we use an entropy-based non-dictionary word segmentation method to generate Tang poetry corpus, which is the basis of domain knowledge extraction. We then propose an active learning mechanism to extract knowledge using machine learning algorithms and crowdsourcing. Domain experts can label or correct extraction results (i.e. entities and their relations) to help train the learning models interactively. Meanwhile, we propose a category-based crowdsourcing mechanism to facilitate domain experts labeling extraction results. The hypothesis is that workers (i.e. domain experts) are specialized in one or more categories of knowledge and can achieve high accuracy. Specifically, we first classify tasks based on categories (e.g. theme, poet), and then assign tasks to workers who have high accuracy on the corresponding categories.

We build a human–computer cooperation and crowdsourcing platform based on CCF. In total, 30 domain professionals, including professionals on DH and Tang poetry, are invited to participate in experiments. Experimental results on Tang poetry data show that CCF can extract high-quality knowledge with good scalability. The extracted knowledge reveals inherent associations among entities of Tang poetry, which form a knowledge graph for global and fine-grained DH studies and intelligent applications.

This paper is organized as follows. In Section 2, we review the related literature. Section 3 provides a framework for this paper. In Sections 4 and 5, we present CCF in detail. Section 6 is the experiment and case study in Tang poetry. Section 7 concludes the paper.

2. Literature review

2.1 Crowdsourcing in digital humanity

Crowdsourcing is an open call for contributions from workers of the crowd to carry out human intelligence tasks (Kazai, 2011). In the era of big data, the popularity and development of the internet have greatly increased the scope and participation of crowdsourcing.

Crowdsourcing is widely accepted as a means for resolving tasks that computers are not good at. Many scholars have explored collecting human intelligence through crowdsourcing projects. Singh *et al.* (2002) built the "Open Mind Common Sense" system to acquire common sense knowledge from the general public. The system supports manual evaluation of crowdsourcing quality. Cristina Sarasua *et al.* (2012) proposed the CROWDMAP model, which can quickly and cost-effectively improve the accuracy of existing ontology alignment schemes. The golden standard is used to assess the accuracy of the crowd-computed results.

Recently, researchers begin to explore crowdsourcing in the cultural heritage domain. Trevor Owens (2013) argued that libraries, archives and museums often invite the public to mark and classify, transcribe, organize and otherwise add value to digital cultural heritage collections. Ridge (2013) proposed that crowdsourcing can help participants to build a deep and valuable connection to cultural heritage through online collaboration in the museum. Carletti *et al.* (2013) explored the correlation between crowdsourcing and library and showed that crowdsourcing is helpful in the generation and management of library resources.

In general, the most typical crowdsourcing initiatives in DH conclude: (1) Correction and Transcription Tasks. The Shanghai Library has published the Sheng Xuanhuai archive transcribing project, which used the collective wisdom to digitize and annotate the Sheng archives (Zhang X *et al.*, 2018). (2) Contextualization. The project "1001 Stories about Denmark" (Yoshimura and Shein, 2011) linked objects that provide contexts such as times, places and personal stories contributed by end-users. (3) Complementing Collection. UK_Soundmap project (British Library, 2010) invited users to provide sound archive and contextual metadata, including a geo-coordinate. (4) Classification. The Powerhouse Museum in Sydney launched a social tagging project to incorporate users' tags in the online catalog (Chan, 2007).

Most of the crowdsourcing initiatives described above can benefit from human inputs. However, humans alone cannot bear the burden of processing the vast textual resources that exist today (Schöch, 2013). Besides, the quality of current cultural heritage crowdsourcing projects relies mainly on manual testing. If there is no proper quality control, crowdsourcing can produce relatively low-quality results (Oomen and Aroyo, 2011).

2.2 Knowledge extraction related to digital humanities

In the field of DH, the need for automated or semi-automated knowledge extraction from large-scale data is widely recognized (Fu *et al.*, 2013). The extracted knowledge mainly includes three elements: entity, relationship and attribute. These elements form a series of high-quality factual expressions, which are understandable for the computer to support advanced knowledge services (Alani *et al.*, 2003).

Researchers first propose rule- and dictionary-based methods for knowledge extraction. Bradley Efron did a statistical analysis of Shakespeare's vocabulary. And, he applied the results to determine whether Shakespeare had written the new poem found in the Bodleian Library (Kolata, 1986). Chen Bingzao analyzed *The Dream of the Red Chamber* from a perspective of word frequency and demonstrated that 120 chapters were written by Cao Xueqin (Ma, 2014). However, such knowledge extraction methods require experts to define extraction rules in advance, which are not feasible in big data scenarios.

To improve the efficiency of knowledge extraction, people began to adopt the method of machine learning. *K*-nearest neighbor (KNN) is a classification algorithm. Each sample can be represented by its nearest *k* neighbors. Classification is performed by measuring the distance between different eigenvalues (Keller, 1985). KNN has good classification efficiency in the scenario of high-dimensional data. Hidden Markov model (HMM) is the simplest dynamic Bayesian network generation model to describe a Markov process with implicit unknown parameters. HMM has been widely used in text recognition, picture recognition, etc.

(Morwal, 2012) Long Short-Term Memory (LSTM) networks is a time-cycle neural network that can remember the state of historical data for a long time and automatically determine the key to the optimal time interval. However, the training of LSTM is costly and complex, which requires a large amount of training corpus (Hammerton, 2003). Conditional random field (CRF) is based on a probability map model that follows the Markov property. CRF is widely applied in part-of-speech tagging, Chinese word segmentation and named entity recognition (Zhao, 2006). However, the training of CRF is also costly and complicated. The active learning mechanism aims to train an accurate prediction model with minimum cost by labeling most informative instances (Sinohara and Miura, 2003).

In the DH field, Tom Horton *et al.* (2006) used the nineteenth-century American novels to realize sentiment classification based on machine learning. Celikyilmaz *et al.* (2010) used the ACTM (actor-topic model) model to explore the social network relationships implied by characters in the nineteenth-century western novels. Caccavale and Søgaard (2019) trained a neural language model to select modernist western poetic entities, based on local context windows. Fang *et al.* (2009) adapted natural language processing (NLP) and corpus analysis techniques to structured imagery analysis in classical Chinese poetry.

Besides, studies tried to invite professionals to participate in information extraction. Plaisant *et al.* (2006) explored the letters of the famous the nineteenth-century American poet Emily Dickinson. Through the combination of automatic classification and manual judgment, the indicators of pornographic features in her poems can be found. Gill (2012) proposed the collaborative intelligence theory. Collaborative intelligence not only underlines the cooperation between human and computer to address cross-disciplinary challenges but also focuses on how to realize positive cooperation among people with different expertise. This theory provides useful guidance for human–machine cooperation projects.

2.3 Tang poetry information resources construction

Different from modern poetry, Tang poetry itself has relatively strict rules on syntax, diction and imagery (Yu-Kung, 1971). Different parts of a Tang poetry vary in syntactic properties (Lee *et al.*, 2017). These characteristics make it inefficient for computers to directly process Tang poetry texts and construct Tang poetry information resources.

In the early stage, electronic document databases have appeared. The whole Tang poetry database (Zhengzhou University, 2008) contains 42,863 poetries in the Tang dynasty. The Chinese Text Project (Sturgeon, 2018) stores related literature in the Tang dynasty. In general, the electronic document databases store unstructured electronic texts and pictures instead of structured knowledge. They only support functions such as browsing and simple keyword matching retrieval. With the development of the technology of structured relational database, many structured subject databases have emerged in DH. The China Biographical Database (CBDB) (Harvard University, 2008) is a typical representative. It contains biographical information about 8,700 poets in the Tang dynasty. Besides, the "China Historical Geographic Information System" (Harvard University, 2001) describes the changes of geographical names, administrative structures over time. Wang (2018) built a chronicle map of literature in Tang and Song dynasties. This knowledge is obtained through manual processing of the literature by experts.

With the development of semantic and artificial intelligence technologies, knowledge bases and knowledge graphs have emerged, such as DBpedia (Lehmann *et al.*, 2015), which also contains data of ancient Chinese literature and historical geographies. The Academic Inheritance Knowledge Graph in the Song dynasty (Peking University, 2018), based on CBDB data, constructs an ontology application of academic inheritance relationship, providing dynamic and visual historical knowledge exploration and discovery. The Garden of Tang Poetry (Beijing Normal University, 2018) uses NLP technologies to mine knowledge in the field of Tang poetry. However, this platform is not designed for domain studies and lacks

important domain knowledge due to its ineffective and inefficient knowledge extraction method. Based on the above literature review, existing works have the following shortcomings:

First, there is no end-to-end knowledge extraction framework designed for DH.

Second, the state-of-the-art knowledge extraction methods for DH data are either low-quality or inefficient in a big data environment.

Third, existing crowdsourcing mechanisms do not consider the uniqueness of DH data, especially Tang poetry data, which is usually domain-specific.

3. Cooperative crowdsourcing framework

Based on the collaborative intelligence theory, we propose the CCF for knowledge extraction from Tang poetry data. CCF has to solve the following research problems:

(1) Fu *et al.* (2013) proposed that feedback to predicting results is conducive to improve the accuracy of machine model training. However, different Tang poetry corpus contributes differently to model training. If training instances are actively selected, can the contribution of training instance be maximized, so that the active feedback to the model be realized?

(2) Branson (2010) found human–computer cooperation can improve the quality of information extraction by machine learning algorithms. In the DH domain, is the human–machine collaborative framework suitable for Tang poetry data processing, specifically combining machine learning with crowdsourcing?

(3) In the crowdsourcing process, Zheng (2016) pointed out that different workers have different labeling accuracy and different tasks in different categories could have different difficulty levels. Can crowdsourcing efficiency be further improved by a category-based crowdsourcing mechanism?

(4) Zeng (2017) presented that the extracted knowledge from big data could promote advancement and change of DH. In the field of Tang poetry, how can the extracted knowledge support global and fine-grained humanity studies?

In CCF, we propose a human–machine cooperation mechanism based on active learning to facilitate domain experts labeling and correcting the extracted knowledge. To further scale up and refine the cooperation, we propose a category-based crowdsourcing mechanism to assign the labeling tasks to the corresponding domain experts. As shown in Figure 1, CCF contains Input Engine, Machine Extraction Engine and Crowdsourcing Engine, which form a

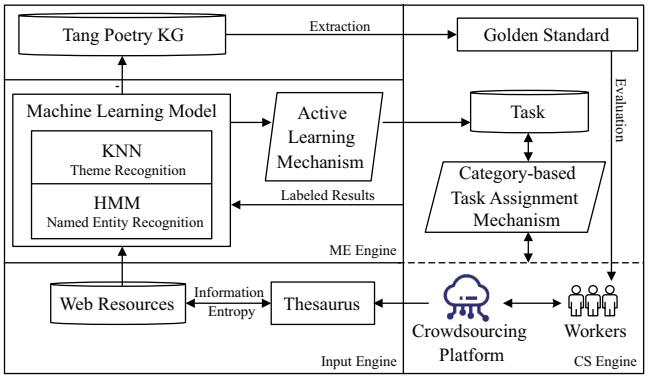

Figure 1.
CCF

two-tier knowledge cycle. The inner cycle is the machine learning model and domain expert interaction, while the outer cycle is the input data and domain expert interaction. Note that domain experts are in the critical position of the framework, as the humanity knowledge is too difficult to be automatically extracted by algorithms without human cooperation.

Input Engine: Because Chinese is written without word delimiters, word segmentation is the key step of processing Chinese texts. However, ancient Tang poetry texts usually have different grammar and writing styles from modern texts, and it is difficult to achieve the desired effect by simply using existing word segmentation tools. Therefore, we design the Input Engine to build the thesaurus of Tang poetry. We use the non-dictionary word segmentation method based on information entropy (Huang, 2003). The thesaurus is integrated with exiting Chinese corpus, including the complete Tang poetry database, CBDB and China Historical Geographic Information System (CHGIS).

Machine Extraction Engine (ME Engine): According to the characteristics of different knowledge extraction tasks, we adopt different machine learning models, including KNN and HMM, to carry out customized knowledge extraction. In addition, we introduce an active learning mechanism to interactively select the most informative instances to provide sufficient information to the machine learning models with good generalization capability.

Crowdsourcing Engine (CS Engine): In ME Engine, the machine learning model will actively generate training corpus set to be labeled. To improve the labeling quality, we crowdsource the labeling tasks to domain experts. The labeled knowledge, in turn, assists the training of the knowledge extraction model in the ME Engine.

Different from the random assignment of tasks, we propose a category-based crowdsourcing mechanism. It supports dynamic worker capacity modeling and category-based task assignment mechanism through the matching between tasks and workers.

4. Cooperative extraction based on active learning mechanism

In this section, we introduce in detail our cooperative knowledge extraction based on active learning, i.e. the ME engine.

4.1 Knowledge extraction model

The theme of Tang poetry represents the thoughts and feelings expressed by poets, which is very important in understanding and appreciating poetries. However, the existing appreciation data are only for a small part of poetries, and some poetries lack relevant analysis of theme. Such data sparsity problem poses challenges for researchers to accurately classify poetries and carry out poetry theme association analysis in global perspectives. Therefore, we take theme recognition as the first step of knowledge extraction.

Secondly, Time, Person, and Location described in the poetries are of great significance to the analysis of background, thoughts, feelings and styles of poetry creation. Meanwhile, these attributes can be applied in some special issues like textual criticism on Tang poetries, historical geography (Zhou, 2007) and lexical grammar evolution analysis. However, it would be very inefficient if these characteristics are extracted manually by experts. Through the human–machine cooperative extraction method, more efficient knowledge extraction can be realized, which helps researchers to conduct a comprehensive and fine-grained analysis of Tang poetry. Therefore, we take the named entities recognition such as time, person and location as the second step of knowledge extraction in Tang poetry.

4.1.1 Theme extraction. The identification of the theme attribute of Tang poetry is equivalent to classifying Tang poetries according to theme attribute. Firstly, we set up five categories of the theme: life experience (LE), homesickness and missing somebody (HM), farewell (FW), concern for the national fate (CN), political intention (PI).

Different from modern Chinese texts, Tang poetry texts are too short to contain enough semantic information of themes. As KNN is suitable for multi-classification problems, especially for the instance sets to be classified, that have a lot of overlap in the category domain. We design a theme classification model based on KNN.

The basic idea of KNN is to find out the nearest k instance points in the training concentration based on some distance measurement when it is given a test instance, and then make a prediction based on the information of the nearest k instances.

We first generate the vector model of poetry through word2vec. Then, we use the thesaurus in Input Engine to conduct word segmentation of Tang poetry texts in the training corpus. After word segmentation, each word inherits the classification label of the original Tang poetry, and all labeled words are converted to quantification. Then, Euclidean distance is used to measure the distance between test data and various training data (the calculation formula is as follows), and the theme category with the highest probability is returned as the prediction result of test data. In the confidence measurement of results, Gaussian weighting is carried out on the Euclidean distance to obtain the confidence of the classification results.

The calculation formula is as follows:

$$f(x) = \alpha e^{-\frac{(x-\beta)^2}{2\gamma^2}} \tag{1}$$

α is the height of the Gaussian curve, β is the offset of the curve centerline on the x-axis and γ is the half-peak width (the width of the distance between peaks of the function).

4.1.2 Named entity recognition. To extract time, person and location attributes in Tang poetry, it is necessary to label the meaning of specific words in Tang poetry. HMM can well capture the characteristic phenomenon and location information of a named entity, and HMM is efficient and easy to train. Therefore, we design a named entity recognition mechanism based on HMM.

In this experiment, the input corpus of HMM is a word segmentation sequence with Tang poetry entities, containing three types of entities (time, person and location). We use BIO encoding in POS tagging. In particular, we denote X as a noun phrase (NP) chunking so that we can define three new tags:

B-np: the beginning of a noun phrase chunk

I-np: inside of a noun phrase chunk

O: outside of a noun phrase chunk

It can be divided into seven categories in specific poetry annotation of Tang dynasty: B-time, I-time, B-person, I-person, B-location, I-location, O. The probability matrix [P_1, P_2, P_3, P_4, P_5, P_6, P_7] of each word is obtained by Viterbi decoding, corresponding to seven annotation categories. As a result, the predicted annotation result is Max(P). The confidence of the result is calculated as follows:

$$\text{Confidence}(h_i) = \frac{\text{Max}(P)}{\text{Sum}(P)} \tag{2}$$

4.2 Active learning mechanism

The training of the knowledge extraction model largely depends on the quality of training corpus. However, different data instances have different training contributions to the model. Compared with random sampling and annotation, the active selection of annotation instances with high contribution can help the machine learning model achieve higher prediction accuracy (Fu *et al.*, 2013). Through the active selection of instances, the most valuable

instances are chosen to be labeled to form the training corpus, to improve the prediction performance of the machine learning model.

Algorithm 1 Active learning mechanism in ME Engine

Input: Initial Labeled Instance Set N_l (n_i, G^t), Unlabeled Instance Set N_u , Model Training Result $R(n_i, s, c)$, Prediction Accuracy of the Model θ

Output: ME Model

1: **while prediction accuracy** ≤ θ **do**
2: *ME model* ← train the model based on N_l;
3: $N_u \leftarrow N \setminus N_l$;
4: **for** each instance n_i in N_u **do**
5: $r_i \leftarrow$ predict n_i based on *ME model*;
6: **end for**
7: $n^* \leftarrow argmin_i(r_i . c)$;
8: $N_u \leftarrow N_u \setminus n^*$;
9: $(n_i, G^t) \leftarrow$ Label n_i in Crowdsourcing Engine;
10: $N_l \leftarrow N_l \cup (n_i, G^t)$;
11: *ME model* ← update based on N_l;
12: **end while**

As shown in Algorithm 1, while the predicting accuracy does not reach the specified threshold θ, the machine learning model ME is trained based on labeled instances set N_l. The set contains instances and corresponding labeled results (Lines1 and 2). Then, unlabeled instance set N_u is obtained by subtracting labeled instance set from training corpus (Line 3). The ME model is used to predict and analyze samples in the unlabeled instance set. The classification results and results confidence of the instances are calculated (Lines 4–6). Lower confidence indicates the model does not have enough knowledge to judge this instance. Conversely, if the instance is added to the labeled instance set, the prediction effect of the model will be improved. Therefore, we conduct a confidence ranking. We select the instances with the lowest confidence (Line 7). These instances are deleted in the unlabeled instance set (Line 8). Then, we publish them to the Crowdsourcing Engine to obtain annotation results G^t. The instance and labeling results (n_i, G^t) are then added to the labeled instance set (Lines 9 and 10). Finally, the ME *model* is iteratively trained with the updated labeled instance set until the prediction accuracy reaches the expected value (Line 11).

5. Category-based crowdsourcing mechanism

In the crowdsourcing process, we assign the training results with low confidence scores to the domain experts, i.e. workers. The workers then label and correct the results manually. We have the following key observations. Different workers have different levels of labeling, which can be measured by accuracy. The accuracy of labeling is higher when workers are familiar with their tasks, and vice versa. Figure 2 shows the workflow of the Crowdsourcing Engine. We evaluate workers in real time by calculating their annotation accuracy in each category of tasks. Based on the evaluation results, we propose a category-based crowdsourcing mechanism to assign a specific category of tasks to high-quality workers with specialized knowledge, to improve the accuracy and efficiency of crowdsourcing. We will introduce the details in the following subsections.

5.1 Task generation

In ME Engine, the active learning mechanism outputs an unlabeled instance set to the crowdsourcing engine. Then, crowdsourcing tasks are generated based on the unlabeled instance set. We set up two forms of tasks:

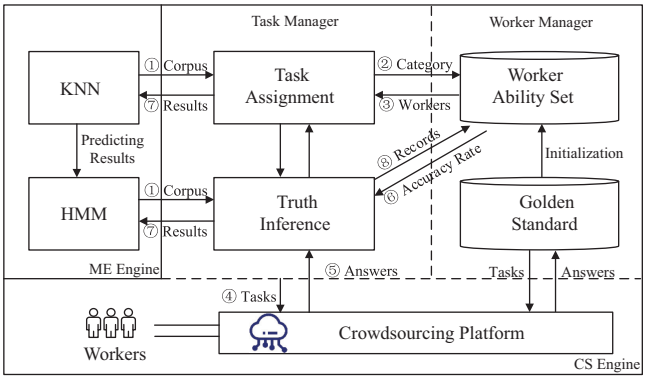

Figure 2.
Workflow of crowdsourcing

(1) Single-question task. This kind of task is for theme recognition. Workers need to choose only one category to which a poetry belongs.

(2) Entity-labeling task. This kind of tasks is for named entity recognition. Workers need to label the words that describe time, person and location attributes in poetries.

As the instance set has been processed by the ME Engine, each instance has a pre-classification, which means that each task t has a category attribute s before labelling.

5.2 Worker ability evaluation

After generating tasks, we need to find a qualified worker for each task. The evaluation of workers' ability is mainly divided into two stages. We first evaluate workers using the golden standard to solve the "cold start problem" and get the initial accuracy of each worker. The golden standard is a set of tasks with the correct answers, which have been labeled by experts. After workers finish golden standard tasks, we calculate each worker's accuracy rates according to Equation (3). Each worker has an accuracy set C, which contains his/her accurate rates in different categories.

$$c_w^s = \frac{E}{E+F} \tag{3}$$

E denotes that the number of tasks in which worker w has labeled correct/confident answers in category s. F denotes the number of tasks in which worker w has not labeled correct/confident answers in category s. c_w^s denotes the accuracy rate of worker w for category s.

If the labeling result is the same as the result of Truth Inference (see section 5.4), it means that the worker hit the task and labeled a confident answer. Otherwise, it means that the worker did not label a confident answer. However, during the labeling process, the ability of workers is not static and may be interfered by internal and external factors. Therefore, in the labeling process, we still need to evaluate the accuracy of workers constantly. Based on workers' labeling records, we use Equation 4 to update workers' accuracy sets.

5.3 Tasks assignment

Before task assignment, we have obtained category attributes of tasks and accuracy rates of workers in each category. Then, we assign the tasks of a specific category to workers who have a high accuracy rate in this category. The specific algorithm is as follows:

Algorithm 2 Category-based task assignment

Input: Tasks set $T(t_i, s)$, Workers set W (w_j, s, c_w^s), Worker ability set $C(s, w_j, c_w^s)$, Assignment Pairs $P(t_i, w_j)$, Answer set $A(t_i, w_j, a_w^t)$
Output: Labeled results
1:**begin**
2: **for** t_i **in** T **do**
3: **for** w_j **in** W **do**
4: **if** $t_i.s == w_j.s$ **then**
4: $w_k \leftarrow$Top-k workers in $w_j.s$ based on C;
5: **end if**
6: $P \leftarrow P(t_i, w_k)$;
7: **end for**
8: $A \leftarrow$ labelled results of P on the crowdsourcing platform;
9: **for** a_w^t **in** A **do**
10: G^t = TruthInference(a_w^t, c_w^s); //G^t: tasks with confident answers
11: **end for**
12: $result \leftarrow result \cup G^t$;
13: **return** *result*
14:**end**

The category-based crowdsourcing mechanism firstly is to match the set of tasks with the set of workers, which means the accuracy of the worker is ranked according to the category attribute s of the task. Then the *top-k* workers in the category s are selected (Lines 1–5). Consequently, pairs of worker and task P are generated and posted to the crowdsourcing platform. Then, the labeling results A can be obtained (Lines 6 and 7). According to the worker's category accuracy, we use the Truth Inference formula (5) to infer the correct answer to the task (Lines 8–10). The tasks with confident answers G^t will be added to labeled results set (Line 11) and be used for training *ME model* in ME engine finally.

5.4 Truth inference

To find the correct answer, which has a high degree of confidence, we ask multiple workers to label the same task and then conduct truth inference. For single-question tasks, we use worker ability set to do weighted calculations. The Truth Inference formula is as follows:

$$G^t = \arg\max_{a \in P} \sum_{w \in W} c_w^s . d\left(a_w^t = a^*\right) \tag{4}$$

where c_w^s denotes the accuracy rate of worker w in category s. a_w^t denotes the answer of worker w to task t. a^* represents a confident answer. $d()$ is an indicator function, which outputs 1 when a_w^t is the same as a^*; otherwise, outputs 0.

For entity labeling tasks, we choose high-frequency words as the final labeling result.

6. Experiment

6.1 Experiment setting

To control the quality of crowdsourcing, we chose professionals from relevant institutions (Li, 2010) as our potential experiment objects. In this experiment, we invited 30 students from Wuhan University, Hunan Normal University, Central China Normal University, Anhui Normal University and other universities. This experiment was conducted online. We built a special WeChat group to inform workers of the basic experiment procedures and provide

them with real-time help during the experiment. And, workers can also use this WeChat group to promptly give feedback and communicate with each other.

We chose the complete Tang poetry database as the training corpus. The complete Tang poetry database contains all the existing Tang poetries. Besides, our experiment was based on the Amazon Mechanical Turk (AMT) platform. AMT is a well-known crowdsourcing platform where requesters can publish the human intelligence tasks (HIT), and workers get rewards by completing tasks (Kittur, 2008). As AMT only supports the random allocation of HITs, we have built an extended task assignment mechanism. Based on this mechanism, we assigned tasks to specific workers and ultimately got their answers (see Figure 3).

Table I shows the statistics of tasks and workers collected from AMT. Before assigning crowdsourcing tasks, the golden standard is assigned to evaluate workers, which contains 15 golden tasks. Then, every 25 tasks are grouped into a batch in a HIT.

6.2 Competing approaches

In this experiment, we designed two different information extraction tasks: theme classification and named entity recognition. We chose different exiting approaches to compare the quality of information extraction. We considered the comparing approach IE: the same machine learning models KNN and HMM are used in CCF without a crowdsourcing mechanism. By comparing CCF with IE, we can verify the validity of the whole CCF.

After verifying the overall performance of CCF, we evaluated the components of CCF: active learning mechanism and category-based crowdsourcing mechanism with the exiting work. We considered three approaches: DOCS: It performs crowdsourcing through domain aware task assignment mechanism (Zheng *et al.*, 2016); however, it does not do any processing on the generation of tasks. Through the comparison of CCF and DOCS, we can verify the effectiveness of active learning mechanism. AL: It takes an active learning mechanism to generate tasks (Culotta and McCallum, 2005), but randomly assign tasks regardless of the characteristics of tasks and workers. By comparing CCF and AL, we can verify the validity of the category-based crowdsourcing mechanism.

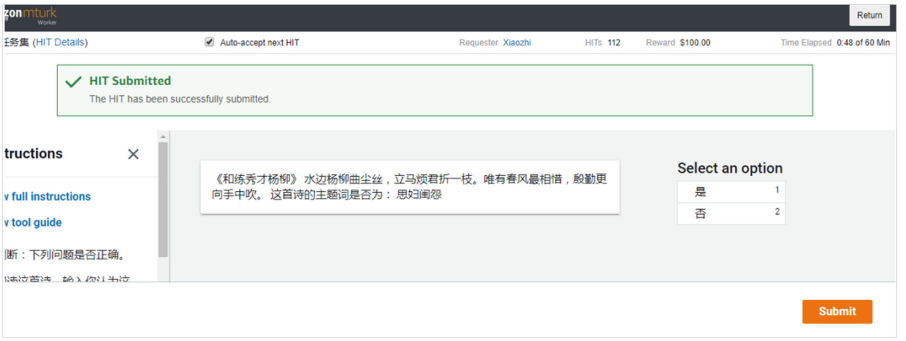

Figure 3.
The interface of crowdsourcing platform

	Tasks	Workers	Workers per HIT	Tasks per HIT
Theme recognition	5,250	30	3	25
Named entity recognition	750	30	1	25

Table I.
Statistics of tasks and workers

We used F1 value as the evaluation metric of knowledge extraction quality. F1 parameters combine the results of accuracy and recall rate. The larger the F1 parameters are, the higher the quality of knowledge extraction is.

$$\mathrm{F1} = 2 * \frac{\mathrm{precision} * \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{5}$$

Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives:

$$\mathrm{Precision} = \frac{\mathrm{True\ \ Positive}}{\mathrm{True\ \ Positive} + \mathrm{False\ \ Positive}} \tag{6}$$

Recall is defined as the number of true positives divided by the total number of elements that belong to the positive class:

$$\mathrm{Recall} = \frac{\mathrm{True\ \ Positive}}{\mathrm{True\ \ Positive} + \mathrm{False\ \ Negative}} \tag{7}$$

6.3 Evaluation on CCF

We first evaluated the overall performance of the CCF. As can be seen from Figure 4, the F1 value of CCF is much higher than that of IE in both two tasks.

In theme recognition tasks, the F1 value of the IE approach in each category is low, especially in the CN category, the F1 value is less than 0.1. This result proves that it is infeasible to use only existing information extraction method to process Tang poetry texts. By contrast, the F1 value in ALL category of CCF is 0.4538; it gained about 24 percent improvement in the average case (the ALL category), and more than 30 percent improvement in some domains (e.g., CN). In named entity identification tasks, the overall F1 value of CCF is 0.4792; it gained about 11 percent improvement in the average case (the ALL category). The main reason is that CCF can combine the scalability of a machine with human intelligence to achieve effective human–machine collaboration for information extraction.

As shown in Figure 5, we can see that CCF outperforms in most of the categories. In comparison between CCF and AL, AL got lower F1 value in every category, especially in the HM and CN categories. It can be found, with a limited number of labeling tasks, active learning mechanism can greatly improve the training efficiency of the model. In the comparison of CCF and DOCS, the overall F1 value of CCF is 0.4538, while the overall F1 value of DOCS is 0.3981. It shows that the category-based crowdsourcing mechanism can assign tasks to workers who are good at it, thus improving the accuracy rate of crowdsourcing and improving the training performance of the model.

6.4 Evaluation of workers

Finally, we calculated the ability set of workers based on experimental records. Firstly, we listed ten workers' accuracy rate for every category. It can be found in Figure 6 that the accuracy rate of workers' labeling in different task categories varies greatly.

Meanwhile, there are also significant differences in the accuracy of each worker in the same task category. For example, the accuracy rate of Worker 2 in the HM category is 0.88, while his/her accuracy rate is 0.5385 in the FW category. In the PI category, the accuracy rate of Worker 1 reaches 0.8214, while that of Worker 8 is 0.3809. These results validate previous experimental hypotheses and show that it is necessary to assign tasks according to the ability of workers and the categories of tasks.

Besides, we also detected changes in accuracy rate during the crowdsourcing process.

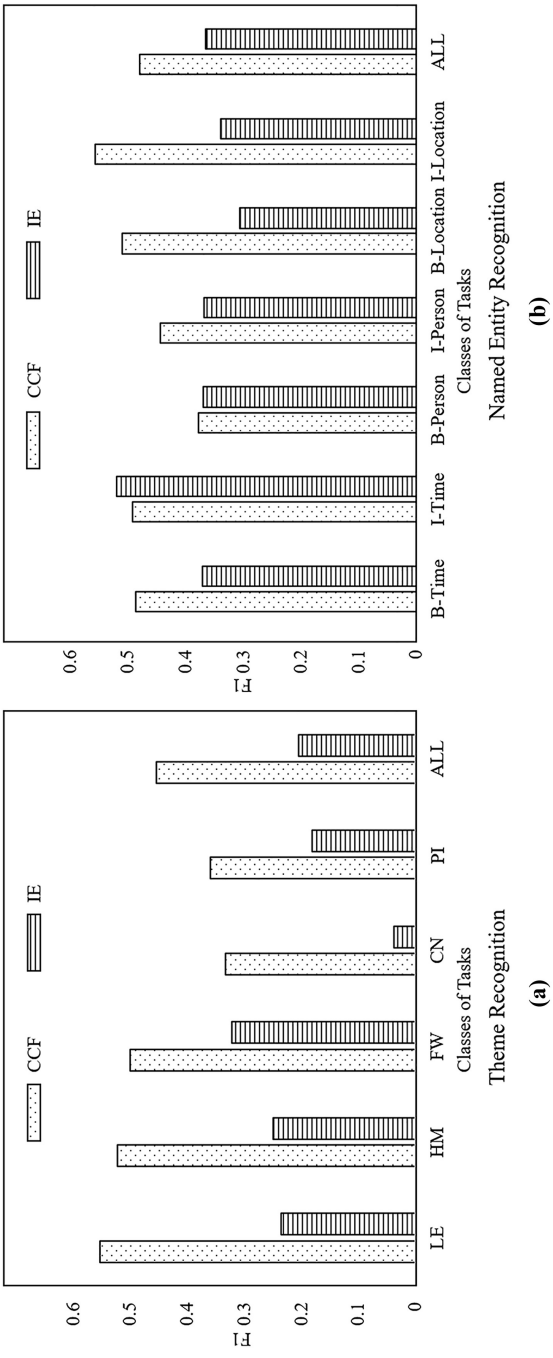

Figure 4.
Evaluation of CCF

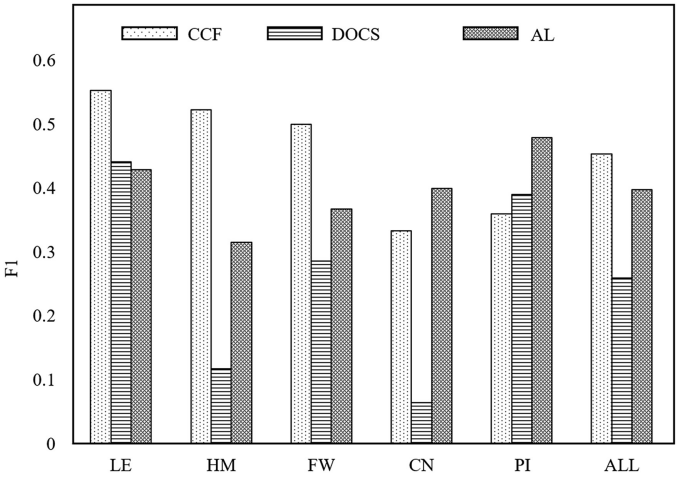

Figure 5.
Evaluation of CCF components

As shown in Figure 7, the accuracy of workers changes when the number of completed tasks increases. In general, workers are more accurate when tagging tasks for one or more particular categories. For example, Worker 1 is good at the HM category, and the accuracy rate is always maintained at 1. Worker 3 is good at FW, LE and HM categories. The accuracy rate of the tasks is maintained above 0.8. However, it is worth noting that workers' performance depends not only on the professional knowledge they have learned. It is also influenced by crowdsourcing environments such as the psychological condition of the workers. Therefore, there may be large fluctuations in the accuracy of the worker during crowdsourcing. For example, the accuracy of Worker 2 changes greatly as tasks increases. For example, the accuracy of Worker 2 in the CN category decreases from 1 to 0.5. This demonstrates that in the crowdsourcing process, it is necessary to update workers' accuracy rate according to labeling results so as to assign them more suitable tasks.

7. Discussion and conclusions

This paper proposes an accurate and extensible CCF based on human–machine collaboration and crowdsourcing. We integrate active learning mechanism with a category-based crowdsourcing mechanism to improve both the accuracy and efficiency of knowledge extraction. Experimental results show that CCF achieves higher accuracy of knowledge extraction than the state-of-the-art methods; the contribution of feedbacks to the training model can be maximized by the active learning mechanism; and our category-based crowdsourcing mechanism can scale up effective human–computer collaboration by considering the specialization of workers in different categories of tasks.

This paper explores the feasibility of applying user-driven collaborative intelligence theory to knowledge extraction in DH. This paper further validates the human-in-the-loop knowledge extraction by integrating active learning and category-based crowdsourcing mechanisms.

CCF extracts knowledge, including entity and entity relations, from multi-source heterogeneous Tang poetry data. The machine-understandable knowledge establishes quantitative associations of entities in different dimensions and provides a global knowledge graph of Tang poetry. These associations reveal hidden knowledge patterns and features, thus achieving intelligent domain knowledge services. Moreover, the knowledge graph can

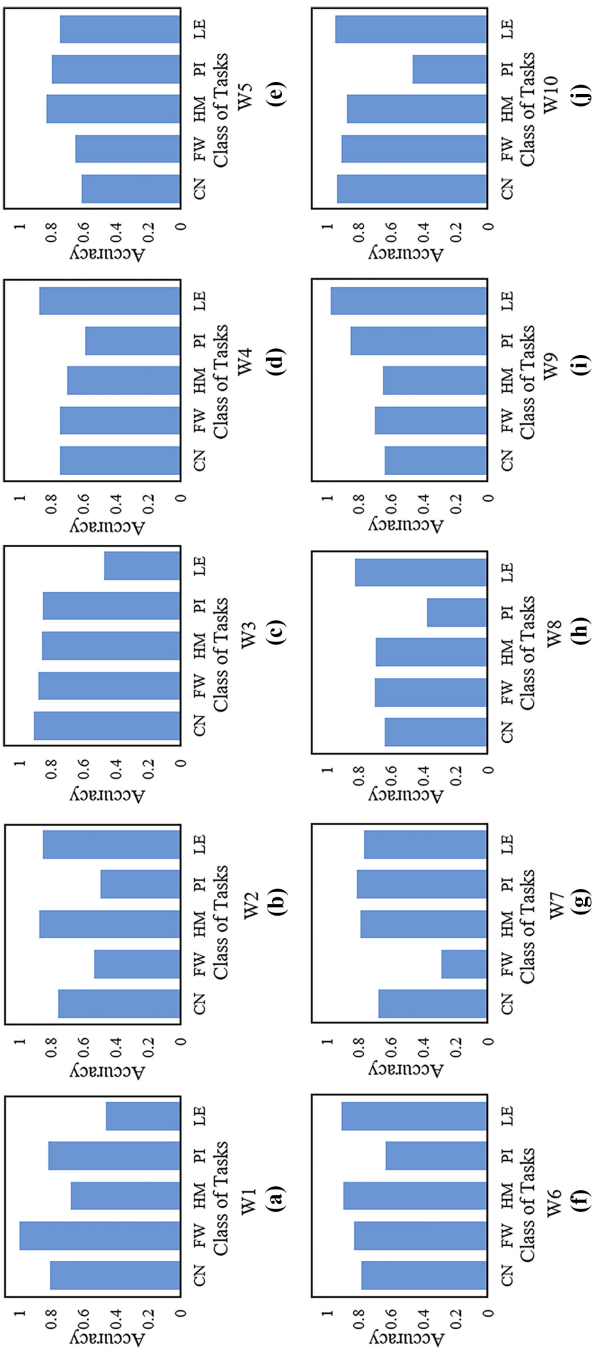

Figure 6.
Accuracy set of workers

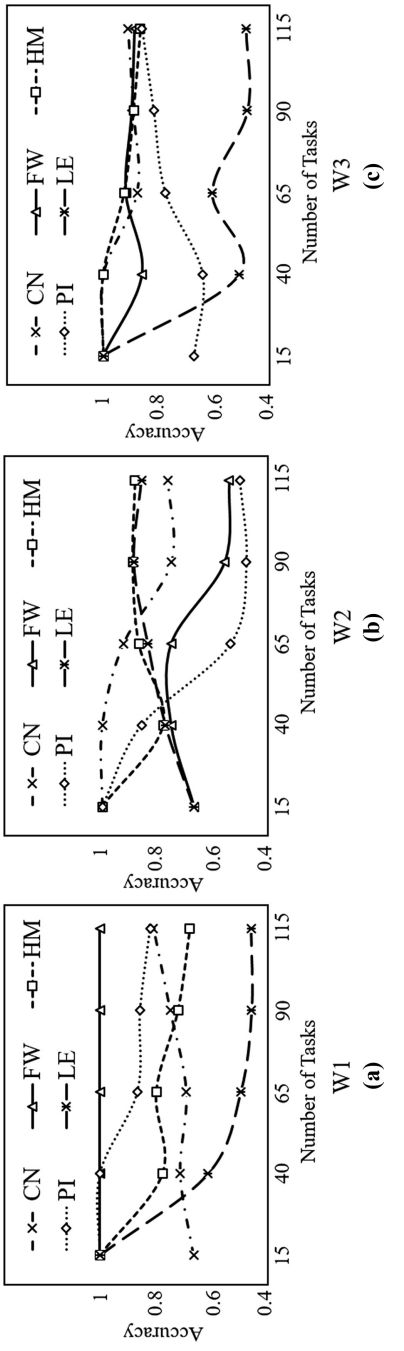

Figure 7.
Changes of workers' accuracy

support humanities research from a quantitative and global perspective, which is different from traditional humanities research.

Because under the CCF, the information extraction algorithm and crowdsourcing task form can be adjusted and determined according to the characteristics of the corpus. For example, when we want to extract knowledge from the Ci poetry of the Song dynasty, we can replace the field thesaurus in Input Engine, task categories and workers in CS Engine and domain-specific knowledge graph with equivalents in the field of Ci poetry of the Song dynasty. Thus, we believe the CCF could be applied to knowledge extraction in other DH fields.

In this paper, we assume that the labeling of workers is accurate, which is not the case in a real-world scenario. In the future, we will take the labeling errors made by domain experts into account and model the transmission and amplification of errors. In addition, we plan to strengthen the theoretical basis of CCF to better reveal the underlying patterns of human–computer collaboration during knowledge extraction.

References

Alani, H., Kim, S., Millard, D.E., Weal, M.J., Hall, W., Lewis, P.H. and Shadbolt, N.R. (2003), "Automatic ontology-based knowledge extraction from web documents", *IEEE Intelligent Systems*, Vol. 18 No. 1, pp. 14-21.

Beijing Normal University (2018), "Garden of Tang poetry", available at: http://poem.studentsystem.org/ (accessed 1 May 2019).

Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P. and Belongie, S. (2010), "Visual recognition with humans in the loop", *European Conference on Computer Vision*, Springer, Berlin, Heidelberg, pp. 438-451.

British Library (2010), "Capturing the sounds of the UK", available at: https://www.bl.uk/press-releases/2010/august/capturing-the-sounds-of-the-uk/ (accessed 9 April 2012).

Caccavale, F. and Søgaard, A. (2019), "Predicting concrete and abstract entities in modern poetry", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 858-864.

Carletti, L., Giannachi, G., Price, D., McAuley, D. and Benford, S. (2013), "Digital humanities and crowdsourcing: an exploration", *Museums and the Web 2013 Conference*, Portland, Oregon.

Celikyilmaz, A., Hakkani-Tur, D., He, H., Kondrak, G. and Barbosa, D. (2010), "The actortopic model for extracting social networks in literary narrative", *NIPS Workshop: Machine Learning for Social Computing*.

Chan, S. (2007), "Tagging and Searching–Serendipity and museum collection databases", *Museums and the Web 2007 Conference*, San Francisco, pp. 87-99.

Culotta, A. and McCallum, A. (2005), "Reducing labeling effort for structured prediction tasks", *AAAI*, Vol. 5, pp. 746-751.

Fang, A.C., Lo, F. and Chinn, C.K. (2009), "Adapting nlp and corpus analysis techniques to structured imagery analysis in classical Chinese poetry", *Proceedings of the Workshop on Adaptation of Language Resources and Technology to New Domains*, Association for Computational Linguistics, pp. 27-34.

Fu, Y., Zhu, X. and Li, B. (2013), "A survey on instance selection for active learning", *Knowledge and Information Systems*, Vol. 35 No. 2, pp. 249-283.

Gill, Z. (2012), "User-driven collaborative intelligence: social networks as crowdsourcing ecosystems", *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, ACM, pp. 161-170.

Hammerton, J. (2003), "Named entity recognition with long short-term memory", *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4*, Association for Computational Linguistics, pp. 172-175.

Harvard University (2001), "China historical geographic information system project history", available at: http://www.fas.harvard.edu/~chgis/ (accessed 25 July 2019).

Harvard University (2008), "China biographical database project history", available at: https://projects.iq.harvard.edu/cbdb (accessed 28 July 2019).

Horton, T., Taylor, K., Yu, B. and Xiang, X. (2006), "Quite right, dear and interesting: seeking the sentimental in nineteenth century American fiction", *Digital Humanities Conference 2006*, Paris, pp. 81-82.

Huang, J.H. and Powers, D. (2003), "Chinese word segmentation based on contextual entropy", *Proceedings of the 17th Pacific Asia Conference on Language*, Information and Computation, pp. 152-158.

Kazai, G. (2011), "In search of quality in crowdsourcing for search engine evaluation", *European Conference on Information Retrieval*, Springer, Berlin, Heidelberg.

Keller, J.M., Gray, M.R. and Givens, J.A. (1985), "A fuzzy k-nearest neighbor algorithm", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 4, pp. 580-585.

Kittur, A., Chi, Ed H. and Suh, B. (2008), "Crowdsourcing user studies with Mechanical Turk", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM.

Kolata, G. (1986), "Shakespeare's new poem: an ode to statistics", *Science*, Vol. 231, pp. 335-337.

Lee, J., Kong, Y.H. and Luo, M. (2017), "Syntactic patterns in classical Chinese poems: a quantitative study", *Digital Scholarship in the Humanities*, Vol. 33 No. 1, pp. 82-95.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S. and Bizer, C. (2015), "DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia", *Semantic Web*, Vol. 6 No. 2, pp. 167-195.

Li, T. (2010), "Review and reflection on the study of literature in the Tang dynasty in the past 30 years", *Journal of Northwest Normal University (Social Science Edition)*, Vol. 47 No. 5, pp. 39-45.

Ma, J., Li, Y. and Teng, G. (2014), "CWAAP: an authorship attribution forensic platform for Chinese web information", *JSW*, Vol. 9 No. 1, pp. 11-19.

Morwal, S., Jahan, N. and Chopra, D. (2012), "Named entity recognition using hidden Markov model (HMM)", *International Journal on Natural Language Computing*, Vol. 1 No. 4, pp. 15-23.

Oomen, J. and Aroyo, L. (2011), "Crowdsourcing in the cultural heritage domain: opportunities and challenges", *Proceedings of the 5th International Conference on Communities and Technologies*, ACM, pp. 138-149.

Owens, T. (2013), "Digital cultural heritage and the crowd", *Curator: The Museum Journal*, Vol. 56 No. 1, pp. 121-130.

Peking University (2018), "Academic inheritance knowledge graph in Song dynasty", available at: http://dh.kvlab.org/cbdb_kg/ (accessed 1 May 2019).

Plaisant, C., Yu, B., Kirschenbaum, M.G., Rose, J., Auvil, L., Smith, M.N., Clement, T. and Lord, G. (2006), "Exploring erotic's in Emily Dickinson's correspondence with text mining and visual interfaces", *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM, pp. 140-151.

Ridge, M. (2013), "From tagging to theorizing: deepening engagement with cultural heritage through crowdsourcing", *Curator: The Museum Journal*, Vol. 56 No. 4, pp. 435-450.

Sarasua, C., Simperl, E. and Noy, N.F. (2012), "Crowdmap: crowdsourcing ontology alignment with microtasks", *International Semantic Web Conference*, Springer, Berlin, Heidelberg, pp. 525-541.

Schöch, C. (2013), "Big? Smart? Clean? Messy? Data in the humanities", *Journal of Digital Humanities*, Vol. 2 No. 3, pp. 2-13.

Singh, P., Lin, T., Mueller, E.T., Lim, G., Perkins, T. and Zhu, W.L. (2002), "Open Mind common sense: knowledge acquisition from the general public", *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*, Springer, Berlin, pp. 1223-1237.

Sinohara, Y. and Miura, T. (2003), "Active feature selection based on a very limited number of entities", *International Symposium on Intelligent Data Analysis*, Springer, Berlin, Heidelberg, pp. 611-622.

Sturgeon, D. (2018), "Unsupervised identification of text reuse in early Chinese literature", *Digital Scholarship in the Humanities*, Vol. 33 No. 3, pp. 670-684.

Wang, Z. (2018), "Chronicle map of literatures in Tang and Song dynasties", available at: https://souyun.cn/PoetLifeMap.aspx (accessed 30 July 2019).

Yoshimura, K. and Shein, C. (2011), "Social metadata for libraries, archives and museums Part 1: site reviews", available at: http://www.oclc.org/research/publications/library/2011/2011-02.pdf/ (accessed 10 November 2019).

Yu-Kung, K. and Tsu-Lin, M. (1971), "Syntax, diction, and imagery in T'ang poetry", *Harvard Journal of Asiatic Studies*, Vol. 31, pp. 49-136.

Zeng, M.L. (2017), "Smart data for digital humanities", *Journal of Data and Information Science*, Vol. 2 No. 1, pp. 1-12.

Zhang, X., Song, S. and Zhao, Y. (2018), "Motivations of volunteers in the Transcribe Sheng project: a grounded theory approach", *Proceedings of the Association for Information Science and Technology*, Vol. 55 No. 1, pp. 951-953.

Zhao, H., Huang, C.N. and Li, M. (2006), "An improved Chinese word segmentation system with conditional random field", *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pp. 162-165.

Zheng, Y., Li, G. and Cheng, R. (2016), "Docs: a domain-aware crowdsourcing system using knowledge bases", *Proceedings of the VLDB Endowment*, Vol. 10 No. 4, pp. 361-372.

Zhengzhou University (2008), "The complete Tang poetry database", available at: http://www3.zzu.edu.cn/qts/ (accessed 1 July 2019).

Zhou, S. and Li, C. (2007), "Tang poetry and historical geography", *Yindu Academic Journal*, No. 1, pp. 61-71.

Corresponding author
Liang Hong can be contacted at: hong@whu.edu.cn