Knowledge Graphs Meet Crowdsourcing: A Brief Survey

Meilin Cao, Jing Zhang(✉), Sunyue Xu, and Zijian Ying

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
1527449673@qq.com, {jzhang,320127010185}@njust.edu.cn, 834391247@qq.com

Abstract. In recent years, as a new solution for hiring laborers to complete tasks, crowdsourcing has received universal concern in both academia and industry, which has been widely used in many IT domains such as machine learning, computer vision, information retrieval, software engineering, and so on. The emergence of crowdsourcing undoubtedly facilitates the Knowledge Graph (KG) technology. As an important branch of artificial intelligence that is recently fast developing, the KG technology usually involves machine intelligence and human intelligence, especially in the creation of knowledge graphs, human participation is indispensable, which provides a good scenario for the application of crowdsourcing. This paper first briefly reviews some basic concepts of knowledge-intensive crowdsourcing and knowledge graphs. Then, it discusses three key issues on knowledge-intensive crowdsourcing from the perspectives of task type, selection of workers, and crowdsourcing processes. Finally, it focuses on the construction of knowledge graphs, introducing innovative applications and methods that utilize crowdsourcing.

Keywords: Crowdsourcing · Human computation · Knowledge graphs · Knowledge mining · Ontology construction

1 Introduction

The concept of crowdsourcing was first proposed by Jeff Howe back in 2006 [1]. He pointed out that crowdsourcing was different from outsourcing. Crowdsourcing is the practice of assigning tasks of an organization or company to a non-specific crowd through network platforms. Crowdsourcing solves problems at a lower cost by tap-ping the potential talents in the crowd. As a novel solution to the acquisition of information and knowledge, crowdsourcing has been widely adopted by many disciplines to facilitate their development, such as business intelligence [2], computer vision [3], software engineering [4], information retrieval [5], machine learning [6], biomedical research [7], health science [8], and so on.

Nowadays, the complex knowledge has be represented as a graphical structure, which is called Knowledge Graphs (KGs). The concept was first put forward by Google in 2012,

L. Qi et al. (Eds.): CloudComp 2020, LNICST 363, pp. 3–17, 2021.
https://doi.org/10.1007/978-3-030-69992-5_1

where the new things were applied to its search engine. The introduction of knowledge graphs strengthened the semantic capability of the search engine, making a query search the contents in the level of domain knowledge instead of simply literally matching the strings in Web databases [9]. Knowledge graphs emphasize entities and their relations rather than strings. Similar to resource pages, knowledge graphs need to be built first, then stored, and finally applied.

The construction of knowledge graphs is inseparable from the participation of humans. Comparing with employing domain experts, introducing crowd workers in the construction and refinement of knowledge graphs is cheaper and fast. However, because of the low quality of non-expert workers, the core problem of crowdsourcing is to optimize the matching of tasks and workers and improve the user experience. Furthermore, when crowdsourcing is applied to the knowledge graph creation, it involves a process of extracting human wisdom. Undoubtedly, knowledge-intensive crowdsourcing may be more complicated, where production and utilization of large-scale knowledge will form an ideal cycle, and human wisdom will continue to promote the operation of this cycle.

In this paper, we first briefly review some basic concepts of crowdsourcing and knowledge graphs, especially focusing on the knowledge-intensive crowdsourcing and the construction of knowledge graphs. Then, we discuss three issues on knowledge-intensive crowdsourcing from the perspectives of task type, se-lection of workers, and crowdsourcing processes. Finally, we review some innovative applications and methods in knowledge graph creation where crowdsourcing was utilized.

2 Basic Concepts of Crowdsourcing and Knowledge Graphs

In this section, we briefly review some basic concepts of crowdsourcing and knowledge graphs.

2.1 Characteristics of Crowdsourcing

As a well-known fact, the definition of crowdsourcing was first proposed in 2006 by Jeff Howe [1], a journalist at the Wired Magazine. However, as early as 2005, a Chinese scholar Feng Liu had created a word “witkey”, standing for “the key of wisdom”, to denote the crowdsourcing business model from the perspective of computer technology [10]. Estellés-Arolas et al. [11] summarized as many as 40 different definitions of crowdsourcing. These definitions describe crowdsourcing from different perspectives. Through the comparison and analysis to these definitions, we can come to some basic characteristics of crowdsourcing:

- Participatory online activities;
- Crowdsourcing tasks usually solve complex problems that are difficult to solve individually;
- Distributed problem-solving mechanism.

According to these characteristics, the definition of crowdsourcing can be as follows: crowdsourcing is a kind of participatory online activity, which solves the task that the

machine intelligence alone is difficult to complete by integrating machines and humans on the Internet.

Knowledge-intensive crowdsourcing is a particular kind of crowdsourcing applications as a bridge between the human brain and machines under the scenarios of exploiting and exploring knowledge. Knowledge-intensive crowdsourcing is recognized as one of the most promising areas of the next generation crowdsourcing, mainly because it plays a key role in today's era of knowledge economy [12]. Knowledge-intensive crowdsourcing has some particular characteristics:

- Diversity of tasks and data. The types of tasks include annotation, classification, ranking, clustering, etc., and the types of data include images, text, structured information, etc. The difficulties of the tasks are also different.
- Diversity of crowd workers. The expertise, educational background, intention, and dedication of crowd workers are different.
- The quality of tasks is difficult to evaluate because of the open nature of crowdsourcing and the absence of ground truth. It is also difficult to measure workers' confidence. Moreover, the cost of evaluation itself is rather high.

The main participants of crowdsourcing include task requesters and task completers (also known as workers). The workflow of task requesters usually includes four steps as follows: 1) Design crowdsourcing tasks; 2) Release the crowdsourcing tasks and wait for the results; 3) Filter the results according to predefined rules for quality control; 4) Integrate results and pay the workers via platforms. The activities of workers include: 1) Select crowdsourcing tasks within their interests and also according to their qualifications; 2) Accept the tasks; 3) Perform the tasks; 4) Submit the answers and get the payments.

2.2 Applications of Knowledge-Intensive Crowdsourcing

During the past decade, researchers and engineers developed various applications that belong to the category of knowledge-intensive crowdsourcing. In this section, we use a few very different examples to illustrate its huge practical value.

- Collaborative editing. As a free and open online encyclopedia, Wikipedia is a well-known crowdsourcing application in the world. It has accumulated more than 3 billion words and numerous knowledge items through crowdsourcing. It will be a huge project to formalize this huge knowledge network into a knowledge graph.
- Urban planning. In urban planning activities home and abroad, the practice of crowdsourcing to encourage public participation has become mature [13]. The Nexthamburg website [14] provides a planning platform for the public in Hamburg, Germany. Each participant participates in the planning and construction of the city's future development by voting, contributing ideas and participating in forums, and further carries out resource crowdfunding on the stadtmacher platform to help realize the public's planning scheme. Similar examples include Mindmixer [15] and Openstreetmap [16] in the United States, and "Zhonggui Wuhan" in China [17] (https://zg.wpdi.cn/).

– Healthcare and Medicine. In recent years, crowdsourcing has also been increasingly used in health science and medical research [18]. Cooper et al. [19] described Foldit, a multiplayer online game, by combining player inputs to determine whether players of the online game MalariaSpot could accurately identify malaria parasites in digitized thick blood smears.
– Marketing. Enterprises can use crowdsourcing to complete marketing related tasks, mainly focusing on product development, advertising and promotion, and marketing research [20]. Dasgupta et al. [21] used a crowdsourcing research website (StreetRx) to solicit data about the price that site visitors paid for diverted prescription opioid analgesics during the first half of 2012. These crowdsourced data provide a valid estimate of the street price of diverted prescription opioids.
– Online learning. In [22], the crowdsourcing method was used to construct Chinese semantic relevance dictionary. Hong et al. [23] combines the incentive mechanism of crowdsourcing with online question-answering technology to simulate teachers' questioning in the real classroom and applies it in MOOC.

2.3 Basic Concept of Knowledge Graphs

The definition of knowledge graph in Wikipedia is as follows: A knowledge graph is the knowledge base that Google uses to enhance its search engine function.[1] In essence, the knowledge graph is a kind of structured semantic knowledge base, which is used to describe concepts and their relationships in the physical world in symbolic form [24]. Knowledge graph is usually designed as a large-scale semantic web, which is composed of entities, concepts and other nodes and attributes, relationships, types and other edges. It is a collection of a large number of triples. Each triplet is composed of subject, predicate, and object.

Triple is the general expression of knowledge graphs [25]. There are four basic types of knowledge tuples:

– <entity, relationship, entity>. E.g., <Mcrosoft, founder, Bill Gates> ;
– <entity, attribute, attribute value >. E.g., <Microsoft, founded time, 1975> ;
– < entity, is-a, concept >. E.g., <Mcrosoft, is-a, listed companies> ;
– <child concept, subclass-of, parent concept>. E.g., <lised company, subclass-of, company>

At present, a number of knowledge graphs have been created for different purposes, such as open domain knowledge graphs (Freebase [26], Dbpedia [27], Wikidata [28], and YAGO[2]), vertical domain knowledge graphs (Linked Life Data[3] and ConceptNet [29]), and Chinese knowledge graphs (Xlore [30] and CN-Dbpedia [31]). Table 1 summarizes the characteristics of some knowledge graphs.

[1] https://en.wikipedia.org/wiki/Knowledge_Graph.
[2] https://yago-knowledge.org/.
[3] https://linkedlifedata.com/.

Table 1. Overview of some popular knowledge graphs.

KG name	Start year	Dependent resources	Scale
ConceptNet	1999	Crowd intelligence	28 million RDF triples
Dbpedia	2007	Wikipedia + Expert knowledge	3 billion RDF triples
YAGO	2007	WordNet + Wikipedia	4,595,906 instances
Freebase	2008	Wikipedia + Domain knowledge + Crowd intelligence	58,726,427 instances
Wikidata	2012	Freebase + Crowd intelligence	42.65 million entries
Xlore	2013	Crowd intelligence	16,284,901 instances

2.4 Construction of Knowledge Graphs

Current construction methods of knowledge graphs are usually based on information extraction in open domains. The construction processes of knowledge graphs typically include three stages [32]: 1) Knowledge extraction, which extracts useful data for business from the original raw data sources; 2) Knowledge fusion, which generally involves knowledge cleaning, entity alignment, and other related processes; 3) Quality evaluation, which judges whether the outcomes meet the predefined requirements. Having a high-quality knowledge graph, we can further carry out knowledge reasoning on it and mine hidden knowledge.

Knowledge Extraction. Knowledge extraction is the primary work of constructing knowledge graphs, including entity extraction, relationship extraction, and attribute extraction [33]. 1) The commonly used entity extraction methods include rule- and dictionary-based, statistical learning-based, and open domain-based extraction methods. Based on the statistics of the characteristics and laws of Chinese place names, Shen et al. [34] sum up the algorithm of Chinese place names, and put forward the reliability probability of word formation and place name continuation to balance the recall and accuracy. Zheng et al. [35] introduced a method of entity recognition based on corpus, extracted and analyzed the frequency of Chinese surname and given name words on the basis of large-scale corpus, and then combined with the rules of context information to determine the place name. Lin et al. [32] proposed a maximum entropy algorithm based on the dictionary, making the recall and accuracy of entity extraction above 70%. Table 2 shows the existing entity extraction methodologies with their advantages and disadvantages. 2) Relation extraction techniques usually can be categorized into template-based methods, lexicon-driven methods, and machine learning-based methods [36]. 3) In terms of attribute extraction, Yang et al. [37] proposed a heuristic attribute extraction method based on rules. Guo et al. [38] used conditional random fields (CRFs) and support vector machines (SVM) to construct collaborative classifiers for attribute and attribute value extraction. In the open test, the accuracy of the collaborative classifier reached 84.4%, and the recall reached 82.7%.

Knowledge Fusion. Knowledge fusion is an important step in the process of knowledge graph construction. After knowledge extraction, the original knowledge can be

Table 2. Entity-extraction methodologies with their advantages and disadvantages.

Methodologies	Advantages	Disadvantages
Rule-based	High accuracy and recall rates can be achieved on small datasets	As the size of a dataset increases, the construction time of the rule set becomes longer and the portability gets worse
Statistical model-based	Little dependence on language and good portability	The correctness of statistical methods and the reliability of statistical sources have a greater impact on the results
Machine (deep) learning-based	Directly take the vector of words in the text as input, without relying on artificially defined features	The implementation techniques are more complicated

obtained. Due to the wide range of knowledge sources in the knowledge map, the quality of knowledge is uneven, the knowledge from different data sources may be repeated, and the correlation between knowledge is not clear enough. Thus, knowledge fusion must be carried out [25]. Knowledge fusion is a high-level abstraction of knowledge organization mode. Key techniques include entity disambiguation, coreference resolution, etc. Entity disambiguation refers to the elimination of different meanings of the same entity. Tan et al. [39] proposed a NED algorithm combining entity linking and entity clustering for entity disambiguation. Ning and Zhang [40] proposed a hierarchical clustering method based on heterogeneous knowledge base to solve the problem of entity disambiguation, and used the Hadoop platform to cluster entity information objects extracted from Wikipedia. Coreference resolution refers to the elimination of the same meaning of different entities. Wang et al. [41] proposed a coreference resolution method based on a decision tree, which combines statistics and rules, and uses rules to filter examples with attribute conflicts. Their method achieved a successful elimination rate of 82.59%. Peng and Yang [42] introduced a maximum entropy model to resolute coreference. By training the model, the problem of common reference resolution is solved, and the improvement is significant.

Knowledge Evaluation. After building the knowledge graph, we need to evaluate the scale and quality of the knowledge map. Mendes et al. [43] proposed a framework for quality assessment, namely Sieve, which has been integrated into the Linked Data Integration Framework (LDIF). Using the Sieve, users can flexibly design their own quality assessment standards. Fader et al. [44] manually annotated entities and relationships in 1000 sentences, and used the results as training sets, and then used the logistics regression model to evaluate the quality of the results. In addition to Sieve, Zaveri et al. [45] also listed dozens of frameworks for knowledge evaluation, and comprehensively reviewed various methods used for quality assessment, and clarified the differences between these methods.

3 Key Issues on Knowledge-Intensive Crowdsourcing

This section briefly reviews three key issues on knowledge-intensive crowdsourcing from the perspectives of tasks, workers, and crowdsourcing processes.

3.1 What Tasks are Suitable for Crowdsourcing

For the purpose of saving budget and time, people generally select the most important tasks or tasks that machines cannot handle but humans are easy to complete to post them on the crowdsourcing platforms. In the process of creating knowledge graphs, crowdsourcing can hand over the task of entity matching and ontology matching. Wang et al. [46] elaborated on the problem of using crowdsourcing for entity matching. Unlike the existing methods (publishing all candidate pairs to the crowdsourcing platform), they studied the relationship between candidate pairs and transferring relationships to reduce overhead. For example, if the entity pairs o_1 and o_2 match, and o_2 and o_3 match, then (o_1, o_3) does not need to be posted on the crowdsourcing platform to make an inference, since the matching of o_1 and o_3 can be obtained automatically. Zhang et al. [47] explored how to use crowdsourcing to reduce the uncertainty of pattern matching (that is, to find the correspondence between the elements of two given patterns). They used probability calculation to locate the correspondence that mostly needs to be determined by crowdsourcing and then judged which group has the highest corresponding probability. Lin et al. [48] studied the application of crowdsourcing in knowledge graph cleaning. They proposed an algorithm to measure which edge to clean would maximize the uncertainty of the system, thereby increasing the time and cost of crowdsourcing. Mo et al. [49] proposed a novel pairwise crowdsourcing model to reduce the uncertainty of top-k ranking using a set of domain experts. For the first k questions and answers based on the knowledge graph, if the given query is compared, the comparison will have a sequential order, which is very vague. At this time, one can make a comparison for it, which is equivalent to a true or false question. Through such short comparisons, the uncertainty of the system can be effectively minimized.

In summary, the selection of crowdsourcing tasks generally follows the principles:

- Preference for small tasks so that workers can use the fragmented time to get paid quickly.
- Local crowdsourcing results will have an impact on the overall situation and this impact needs to be quantified and different tasks have different effects.

3.2 Who Completes Crowdsourcing Tasks

Passive crowdsourcing refers to the mode that when workers actively choose crowdsourcing tasks, and workers may participate in training before performing tasks. The principles of active crowdsourcing task allocation are as follows: randomly assigning tasks, assigning tasks according to worker quality or other criteria (for example, selecting the workers with the highest quality, selecting the nearest workers, or selecting the workers with the closest expected results, etc.). Mo et al. [49] studied a cross-task crowdsourcing problem. That is, the actual labels of data provided by different crowdsourcing

workers in a crowdsourcing environment may be sparse, noisy, and unreliable. They used domain similarity and transfer learning to transfer users' domain skills in reasoning. Zheng et al. [50] introduced a field-based matching method, which decomposed all tasks into 13 fields and calculated the correlation between workers and tasks in each field. Mavridis et al. [51] adopted a skill tree-based matching method, through fine modeling of tasks and participants, and the distance on the tree to represent their correlation. Some tasks may not be able to model the task using only decision trees. At this time, tree-graph combination could be used.

3.3 How to Complete Crowdsourcing Tasks

The good completion of crowdsourcing tasks depends on various factors, mainly including how to design crowdsourcing tasks, how to motivate workers, and how to control the quality of tasks and results.

There are two ways of designing crowdsourcing tasks—explicit crowdsourcing and implicit crowdsourcing. As their names suggest, explicit crowdsourcing means that workers clearly know that they are completing crowdsourcing tasks, while implicit crowdsourcing means that workers do not know the existence of crowdsourcing tasks. An implicit crowdsourcing task is generally hidden behind some other tasks. The facial tasks are used to attract workers, and the workers unconsciously complete the crowdsourcing tasks when completing the facial ones. Compared with explicit crowdsourcing, the implicit scheme has a lower cost and better results. Two different crowdsourcing schemes have different task-designing principles. For explicit crowdsourcing, the designed tasks should be as concise as possible so that they can easily attract many workers and do not require too much completion time. Therefore, the traditional design principle tries to design small tasks. For example, binary-choice (true or false) questions are better than multiple-choice ones, and multiple-choice questions are better than fill-in-the-blank ones, which means the less interaction the better. Also, the UI design of explicit crowdsourcing should be vivid and concise. For implicit crowdsourcing, the crowdsourcing tasks need to be plunged into the facial tasks on the premise that the facial tasks are attractive enough to workers. For example, one can hide a crowdsourcing task in a game and obtains some common-sense knowledge and location information through workers' feedback. The location information can also use the worker's psychological characteristics to arouse their curiosity or distract them so that the workers can complete the crowdsourcing tasks unknowingly. Von Ahn et al. [52] introduced a game called Verbosity. The roles of the game are divided into narrator and guesser. A narrator uses a non-secret word completion template to ask a guesser to guess the secret word. Through this interesting game, users can unknowingly provide common-sense knowledge while enjoying the game (such as true statements like "snow is white"). Ni et al. [53] proposed an alternative ground truth to the eye fixation map in visual attention study called Touch Saliency, which judges the focus of a picture by the position where the user clicks on the screen when seeing the picture. The principles of implicit crowdsourcing task design include: Propose tasks unconsciously; Users can become workers; The facial task meets the needs of users, and the behind one is the crowdsourcing task; The facial task must be attractive enough to users.

Obviously, there have been several approaches to motivate workers, including money, happiness, social influence (can be divided into strong connections such as in WeChat or Facebook, and weak connections such as in Baidu Tieba), and so on. One can use a hybrid incentive mechanism such as using strong social media for publicity at the beginning of the task, using weak social media and monetary incentives after gathering certain popularity, and again using strong social media and monetary incentives to attract remaining workers at the end.

The quality control of crowdsourcing requires to consider correctness, coverage, timeliness, and consistency. Because the quality of crowdsourcing workers is uneven, there may be malicious workers such as fake qualified workers, quick deceivers (aiming to get paid by answering questions indiscriminately), etc. To deal with this tricky situation, the methods such as burying mines (that is, inserting some tasks whose answers are known to check the quality of workers) and backtracking questions (asking questions related to the previous question to prevent users from answering the questions indiscriminately) have been widely adopted. After crowdsourcing tasks are completed, the collected answers need to be verified for credibility. Some easiest method was to use gold standard data to evaluate the quality of workers' outcomes [54]. Using the data with standard answers, the quality of workers can be determined by comparing the results submitted with the standard answers. Another way that does not rely on the gold answers is to use a repeated-answering scheme, where multiple workers independently answer the same questions. The final answers are inferred from the collected multiple noisy answers, which is call the truth inference. During the past decade, a large number of true inference algorithms for crowdsourcing were proposed and achieved good performance [55].

4 Construction of Knowledge Graphs Using Crowdsourcing

In knowledge-intensive crowdsourcing, knowledge graphs can be exploited and explored. This section further focuses on the construction of knowledge graphs, where crowdsourcing provides an effective solution to gather a large amount of knowledge.

4.1 Ontology Construction

Ontology is also called entity in knowledge graphs. This concept originated from western philosophy and describes the objective existence of things. In 1993, Gruber [56] defined ontology as a conceptual and precise specification. In 1998, Studer et al. further extended the concept of ontology and defined it as a clear formal specification of a shared conceptual model [57]. In short, ontology is a data set that is used to describe a domain and the skeleton of the knowledge base. A knowledge graph model needs the support of ontology, and the concept of ontology has been widely concerned in the field of information science in recent years [58]. An ontology consists of five basic elements, including class or concept, relation, function, axiom, and instance. There are usually three approaches for ontology construction—manual construction, automatic construction, and semi-automatic construction.

Here, we introduce semi-automatic ontology construction, which is between manual construction and automatic construction. Because of the high technical requirements of fully automated ontology construction, it is difficult to achieve in most application fields. Therefore, the construction of ontology usually needs human participation, and crowdsourcing platforms just provide a good solution. DiFranzo and Hendler [59] introduced OntoPronto, which is a premeditated game. In this game, two players try to map randomly chosen Wikipedia articles to the most specific classes of the Proton ontology. If they agree on a Proton class for their articles, they will obtain points and proceed to the next specific level. Acosta et al. [60] proposed CrowdSPARQL, a new SPARQL query answer method, which combines machine-driven and human-driven capabilities. When an SPARQL query fails to respond, it will be redirected to the MTurk platform to obtain knowledge. Niepert et al. [61] proposed INPHO, a system combining statistical text processing, information extraction, human expert feedback, and logic programming, which is used to fill and expand the Dynamic Ontology in the field of Philosophy. The system uses crowdsourcing to complete the construction of the concept system.

4.2 Knowledge Mining and Filling

Knowledge is undoubtedly an indispensable part of knowledge graphs. Compared with machines, human beings have inherent advantages in knowledge acquisition and mining, which are shown in three aspects. First, humans can quickly and accurately extract triples from natural language. Second, they can accurately align triple in heterogeneous data sources. Finally, they are good at using common sense database. However, if all of the above work is completed manually, the cost of time and money will be unacceptable. Therefore, in terms of knowledge acquisition, the combination of humans and machines is the mainstream.

Crowdsourcing is widely used for knowledge acquisition, that is, extracting entities and triples from natural languages. Kondreddi et al. [62] proposed a system architecture called Higgins, which shows how to effectively integrate Information Extraction (IE) engine and Human Computing (HC) engine. The system allows players to select or fill in the subject-relation-object triples by setting game problems. Higgins system combines information extraction and human computing, which can be used to edit the relationship between characters in movies or books. First, it creates meaningful questions and answers through the IE engine, and then feeds them into the HC engine for crowdsourcing annotation, which greatly improves the accuracy and reduces the cost.

Crowdsourcing has been used to align entities from heterogeneous knowledge sources. Entity alignment includes entity disambiguation and coreference disambiguation. Zhuang et al. [63] put forward a method of human–computer combination named HIKE for entity alignment. Firstly, rough entity alignment is conducted for knowledge base through machine learning method, and then matched pairs and unmatched pairs are put into crowdsourcing platforms respectively for crowd workers to judge their correctness.

The entity collection based on crowdsourcing aims to collect a large number of open entities. It is also an important application of crowdsourcing in knowledge mining. There may be some challenges such as repetition, omission, and errors when using the method of purely manual entity collection. Crowdsourcing-based entity collection can make

up for the deficiency of purely manual entity collection. Chai et al. [64] proposed an incentive-based crowdsourcing entity collection framework Crowdec, which encourages employees to use incentive strategies to provide more different projects, uses the pricing principle to encourage workers to provide non-repetitive answers, and adopts the worker elimination method to prevent the inefficient use of workers.

4.3 Refinement of Knowledge Graphs

Refinement is one of the important issues in knowledge graph research. Its main tasks include knowledge graph completion and error detection. Here, we introduce the application of crowdsourcing in the refinement stage of knowledge graphs. There are three reasons for introducing crowdsourcing to refine knowledge graphs. First, it is difficult for automatic methods to achieve both high accuracy and wide coverage. Second, documents in the network have a long tail effect, which means a large amount of knowledge is distributed sparsely. Finally, automatic processing technology usually has some defects such as high noise and difficulty in guaranteeing knowledge accuracy.

The function of crowdsourcing to fill the gaps is reflected in all encyclopedia websites, such as Wikipedia. Wikipedia is based on the principle that everyone can participate in it. It has the characteristics of open sharing, interactive collaboration, and comprehensive and accurate information. Wikipedia has also launched its own unique management and editing techniques such as page locking, when the editing level of some main pages reaches the Wikipedia standard, the page is locked to prevent other users from editing at will [65]. Singh et al. [66] proposed a knowledge acquisition system called Open Mind Common Sense, which allows participants to construct and fill natural knowledge templates to obtain facts and common-sense knowledge. In addition, there is crowdsourcing verification based on link prediction and the filling of the domain knowledge graph based on crowdsourcing.

Crowdsourcing error correction usually forms into two trains of thought. In the first scheme, we disclose all data and use crowdsourcing to find and correct errors. This scheme is suitable for large-scale websites with a huge volume of traffic, such as Google. In the second scheme, the machine first locates the possible error locations and then submits them to crowdsourcing. This scheme is suitable for small websites. In fact, as early as 2011 in China, Sogou Maps launched high-speed charging calculation and error correction functions, encouraging users to share charging data to improve the charging function. Currently, various map software also obtains user information through crowdsourcing, encouraging the users to contribute Point-Of-Interest information or correct the routes. Pavlick et al. [67] introduced the application of crowdsourcing in the problem of grammatical error correction, discarding the traditional majority voting method and using crowdsourcing to ensure quality and produce ideal results.

5 Conclusion

Crowdsourcing has provided a good venue for both creating and exploiting knowledge graphs. This paper briefly reviews the recent progress in the intersection of knowledge

graphs and crowdsourcing. The paper first summarizes the characteristics of knowledge-intensive crowdsourcing and the construction procedure of knowledge graphs. Then, it discusses three key issues on knowledge-intensive crowdsourcing from the perspectives of tasks, workers, and crowdsourcing processes, which shapes the contour of current research in this field. Finally, it reviews some innovative applications and methods where crowdsourcing was utilized in the construction of knowledge graphs. We believe that this direction will continue to be a research hot spot in the future.

Acknowledgements. This work has been supported by the National Natural Science Foundation of China under grants 62076130 and 91846104, and the National Key Research and Development Program of China under grant 2018AAA0102002.

References

1. Howe, J.: The rise of crowdsourcing. Wired Mag. **14**(6), 1–4 (2006)
2. Guo, B., Liu, Y., Ouyang, Y., Zheng, V.W., Zhang, D., Yu, Z.: Harnessing the power of the general public for crowdsourced business intelligence: a survey. IEEE Access **7**, 26606–26630 (2019)
3. Kovashka, A., Russakovsky, O., Fei-Fei, L., Grauman, K.: Crowdsourcing in computer vision. Found. Trends® Comput. Graph. Vis. **10**(3), 177–243 (2016)
4. Mao, K., Capra, L., Harman, M., Jia, Y.: A survey of the use of crowdsourcing in software engineering. J. Syst. Softw. **126**, 57–84 (2017)
5. Alonso, O., Lease, M.: Crowdsourcing for information retrieval: principles, methods, and applications. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1299–1300 (2011)
6. Zhang, J., Wu, X., Sheng, V.S.: Learning from crowdsourced labeled data: a survey. Artif. Intell. Rev. **46**(4), 543–576 (2016). https://doi.org/10.1007/s10462-016-9491-9
7. Saez-Rodriguez, J., et al.: Crowdsourcing biomedical research: leveraging communities as innovation engines. Nat. Rev. Genet. **17**(8), 470 (2016)
8. Wazny, K.: Applications of crowdsourcing in health: an overview. J. Glob. Health **8**(1), 010502 (2018)
9. Gomez-Perez, J.M., Pan, J.Z., Vetere, G., Wu, H.: Enterprise knowledge graph: an introduction. In: Pan, J., Vetere, G., Gomez-Perez, J., Wu, H. (eds.) Exploiting linked data and knowledge graphs in large organisations, pp. 1–14. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-45654-6_1
10. Lin, S.F., Lin, F.: A research on the definitions and models of crowdsourcing and their future development. Sci. Technol. Manage. Res. **35**(4), 212–217 (2015)
11. Estellés-Arolas, E., González-Ladrón-De-Guevara, F.: Towards an integrated crowdsourcing definition. J. Inf. Sci. **38**(2), 189–200 (2012)
12. Roy, S.B., Lykourentzou, I., Thirumuruganathan, S., Amer-Yahia, S., Das, G.: Task assignment optimization in knowledge-intensive crowdsourcing. VLDB J. **24**(4), 467–491 (2015). https://doi.org/10.1007/s00778-015-0385-2
13. Wan, Y.L., Xu, Q.W., Liao, P.C., Li, X.X.: Application and prospect of crowdsourcing in urban planning. J. Tsinghua Univ. (Sci. Technol.) **59**(5), 409–416 (2019)
14. Weninger, B., Poplin, A.K., Petrin, J.: Developing a typology of public participation 2.0 users: an example of Nexthamburg.de. In: Proceedings of 15th International Conference on Urban Planning and Regional Development in the Information Society, pp. 191–199 (2010)

15. Nguyen, C., Tahmasbi, N., De Vreede, T., De Vreede, G.J., Oh, O., Reiter-Palmon, R.: Participant engagement in community crowdsourcing. In: European Conference on Information Systems, vol. 41 (2015)
16. Haklay, M., Weber, P.: OpenStreetMap: user-generated street maps. IEEE Pervasive Comput. **7**(4), 12–18 (2008)
17. Xiong, W., Zhou, B.: Construction and thinking of "Zhonggui Wuhan" open platform. Beijing Plann. Constr. **1**, 100–102 (2016)
18. Alialy, R., et al.: A review on the applications of crowdsourcing in human pathology. J. Pathol. Inf. **9**, 2 (2018)
19. Cooper, S., et al.: Predicting protein structures with a multiplayer online game. Nature **466**(7307), 756–760 (2010)
20. Whitla, P.: Crowdsourcing and its application in marketing activities. Contemp. Manage. Res. **5**(1), 15–28 (2009)
21. Dasgupta, N., et al.: Crowdsourcing black market prices for prescription opioids. J. Med. Internet Res. **15**(8), e178 (2013)
22. Ding, Y., Che, W.X., Liu, T., Zhang, M.S.: Constructing word association network by crowdsourcing. J. Chin. Inf. Process. **27**(3), 100–107 (2013)
23. Hong, L., Ran, C.J., Yu, Q.: Implementation of MOOC online question and answer system with crowdsourcing. Libr. Inf. Serv. **58**(19), 118–123 (2014)
24. Qiao, L., Yang, L., Hong, D., Yao, L., Zhiguang, Q.: Knowledge graph construction techniques. J. Comput. Res. Dev. **53**(3), 582–600 (2016)
25. Xu, Z.L., Sheng, Y.P., He, L.R., Wang, Y.F.: Review on knowledge graph techniques. J. Univ. Electron. Sci. Technol. China **45**(4), 589–606 (2016)
26. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 1247–1250 (2008)
27. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) The Semantic Web. ISWC 2007, ASWC 2007. Lecture Notes in Computer Science, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52
28. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)
29. Liu, H., Singh, P.: ConceptNet—a practical commonsense reasoning tool-kit. BT Technol. J. **22**(4), 211–226 (2004). https://doi.org/10.1023/B:BTTJ.0000047600.45421.6d
30. Wang, Z., et al.: XLore: a large-scale English-Chinese bilingual knowledge graph. In: International Semantic Web Conference (Posters & Demos), vol. 1035, pp. 121–124 (2013)
31. Xu, B., et al.: CN-DBpedia: a never-ending Chinese knowledge extraction system. In: Benferhat, S., Tabia, K., Ali, M. (eds.) Advances in Artificial Intelligence: From Theory to Practice. IEA/AIE 2017. Lecture Notes in Computer Science, vol. 10351, pp. 428–438. Springer, Cham. https://doi.org/10.1007/978-3-319-60045-1_44
32. Lin, Y.F., et al.: A maximum entropy approach to biomedical named entity recognition. In: Proceedings of the 4th International Conference on Data Mining in Bioinformatics, pp. 56–61 (2004)
33. Zhang, Z.X., Wu, Z.X., Liu, J.H., Xu, J., Hong, N., Zhao, Q.: Analysis of state-of-the-art knowledge extraction technologies. Data Anal. Knowl. Discov. **24**(8), 2–11 (2008)
34. Shen, D.Y., Sun, M.S., Huang, C.N.: Identifying Chinese Place Names in Unrestricted Text, pp. 68–74. Tsinghua University Press, Beijing (1995)
35. Zheng, J.H., Li, X.: Research on Chinese name recognition method based on corpus. J. Chin. Inf. Process. **14**(1), 7–12 (2000)
36. Xu, J., Zhang, Z.X., Wu, Z.X.: Review on techniques of entity relation extraction. Data Anal. Knowl. Discov. **24**(8), 18–23 (2008)

37. Yang, B., Cai, D.F., Yang, H.: Progress in open information extraction. J. Chin. Inf. Process. **20**(2), 123–151 (2014)
38. Guo, J.Y., Li, Z., Yu, Z.T.: Extraction and relation prediction of domain ontology concept instance, attribute, and attribute value. J. Nanjing Univ. **48**(4), 383–389 (2012)
39. Tan, Y.M., Yang, X.: An named entity disambiguation algorithm combining entity linking and entity clustering. J. Beijing Univ. Posts Telecommun. **37**(5), 36–40 (2014)
40. Ning, B., Zhang, F.: Named entity disambiguation based on heterogeneous knowledge base. J. Xi'an Univ. Posts Telecommun. **19**(4), 70–76 (2014)
41. Wang, Z.Q., Li, L., Wang, C.: Chinese pronominal coreference resolution based on decision tree. J. Beijing Univ. Posts and Telecommun. **29**(4), 1–5 (2006)
42. Pang, N., Yang, E.H.: The research on coreference resolution based on maximum entropy model. J. Chin. Inf. Process. **22**(2), 24–27 (2008)
43. Mendes, P.N., Mühleisen, H., Bizer, C.: Sieve: linked data quality assessment and fusion. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops, pp. 116–123 (2012)
44. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1535–1545 (2011)
45. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semant. Web **7**(1), 63–93 (2016)
46. Wang, J., Li, G., Kraska, T., Franklin, M. J., Feng, J.: Leveraging transitive relations for crowdsourced joins. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, pp. 229–240 (2013)
47. Zhang, C.J., Chen, L., Jagadish, H.V., Cao, C.C.: Reducing uncertainty of schema matching via crowdsourcing. Proc. VLDB Endow. **6**(9), 757–768 (2013)
48. Lin, X., Xu, J., Hu, H., Fan, Z.: Reducing uncertainty of probabilistic top-k ranking via pairwise crowdsourcing. IEEE Trans. Knowl. Data Eng. **29**(10), 2290–2303 (2017)
49. Mo, K., Zhong, E., Yang, Q.: Cross-task crowdsourcing. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 677–685 (2013)
50. Zheng, Y., Li, G., Cheng, R.: DOCS: a domain-aware crowdsourcing system using knowledge bases. Proc. VLDB Endow. **10**(4), 361–372 (2016)
51. Mavridis, P., Gross-Amblard, D., Miklós, Z.: Using hierarchical skills for optimized task assignment in knowledge-intensive crowdsourcing. In: Proceedings of the 25th International Conference on World Wide Web, pp. 843–853 (2016)
52. Von Ahn, L., Kedia, M., Blum, M.: Verbosity: a game for collecting common-sense facts. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 75–78 (2006)
53. Ni, B., et al.: Touch saliency: characteristics and prediction. IEEE Trans. Multimedia **16**(6), 1779–1791 (2014)
54. Zhang, Z.Q., Pang, J.S., Xie, X.Q., Zhou, Y.: Research on crowdsourcing quality control strategies and evaluation algorithm. Chin. J. Comput. **36**(8), 1636–1649 (2013)
55. Sheng, V.S., Zhang, J.: Machine learning with crowdsourcing: a brief summary of the past research and future directions. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 9837–9843 (2019)
56. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing? Int. J. Hum. Comput. Stud. **43**(5–6), 907–928 (1995)
57. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge engineering: principles and methods. Data Knowl. Eng. **25**(1–2), 161–197 (1998)
58. Li, S.P., Yin, Q.W., Hu, Y.J., Guo, M., Fu, X.J.: Overview of researches on the ontology. Comput. Res. Dev. **41**(7), 1041–1052 (2004)

59. DiFranzo, D., Hendler, J.: The semantic web and the next generation of human computation. In: Michelucci, P. (ed.) Handbook of Human Computation, pp. 523–530. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-8806-4_39
60. Acosta, M., Simperl, E., Flöck, F., Norton, B.: A SPARQL engine for crowdsourcing query processing using microtasks–Technical report. Institute AIFB, KIT, Karlsruhe (2012)
61. Niepert, M., Buckner, C., Murdock, J., Allen, C.: InPhO: a system for collaboratively populating and extending a dynamic ontology. In: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 429–429 (2008)
62. Kondreddi, S.K., Triantafillou, P., Weikum, G.: Combining information extraction and human computing for crowdsourced knowledge acquisition. In: 2014 IEEE 30th International Conference on Data Engineering, pp. 988–999 (2014)
63. Zhuang, Y., Li, G., Zhong, Z., Feng, J.: Hike: a hybrid human-machine method for entity alignment in large-scale knowledge bases. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1917–1926, November, 2017
64. Chai, C., Fan, J., Li, G.: Incentive-based entity collection using crowdsourcing. In: 2018 IEEE 34th International Conference on Data Engineering (ICDE), pp. 341–352 (2018)
65. Zhao, F., et al.: Research progress on Wikipedia. Doctoral dissertation. J. Univ. Electron. Sci. Technol. China **39**(3), 321–334 (2010)
66. Singh, P., Lin, T., Mueller, E.T., Lim, G., Perkins, T., Li Zhu, W.: Open mind common sense: knowledge acquisition from the general public. In: Meersman, R., Tari, Z. (eds.) On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE. OTM 2002. Lecture Notes in Computer Science, vol. 2519, pp. 1223–1237. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-36124-3_77
67. Pavlick, E., Yan, R., Callison-Burch, C.: Crowdsourcing for grammatical error correction. In: Proceedings of the Companion Publication of the 17th ACM conference on Computer Supported Cooperative Work & Social Computing, pp. 209–212 (2014)