2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing

A Survey of Crowdsourcing Systems

Man-Ching Yuen[1], Irwin King[1,2], and Kwong-Sak Leung[1]
[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
[2]AT&T Labs Research, San Francisco, USA
{mcyuen, king, ksleung}@cse.cuhk.edu.hk; irwin@research.att.com

***Abstract*—Crowdsourcing is evolving as a distributed problem-solving and business production model in recent years. In crowdsourcing paradigm, tasks are distributed to networked people to complete such that a company's production cost can be greatly reduced. In 2003, Luis von Ahn and his colleagues pioneered the concept of "human computation", which utilizes human abilities to perform computation tasks that are difficult for computers to process. Later, the term "crowdsourcing" was coined by Jeff Howe in 2006. Since then, a lot of work in crowdsourcing has focused on different aspects of crowdsourcing, such as computational techniques and performance analysis. In this paper, we give a survey on the literature on crowdsourcing which are categorized according to their applications, algorithms, performances and datasets. This paper provides a structured view of the research on crowdsourcing to date.**

***Index Terms*—crowdsourcing; survey**

I. Introduction

Nowadays, many tasks that are trivial for humans continue to challenge even the most sophisticated computer programs, such as image annotation. These tasks cannot be computerized. Prior to the introduction of the concept of crowdsourcing, traditional approaches for solving problems that are diffcult for computers but trivial for humans focused on assigning tasks to employees in a company. However, it increases a company's production costs.

To reduce a company's production costs and make more efficient use of labor and resources, crowdsourcing was proposed [33]. Crowdsourcing is a distributed problem-solving and business production model. In an article for Wired magazine in 2006, Jeff Howe defined "crowdsourcing" as "an idea of outsourcing a task that is traditionally performed by an employee to a large group of people in the form of an open call" [32]. An example of crowdsourcing tasks is the creative drawings, such as the Sheep Market [3], [46]. The Sheep Market is a web-based artwork to implicate thousands of workers in the creation of a massive database of drawings. Workers create their version of "a sheep facing to the left" using simple drawing tools. Each worker is responsible for a drawing receives a payment of two cents for his labor.

The explosive growth and widespread accessibility of the Internet have led to surge of research activity in crowdsourcing. Currently, all crowdsourcing applications have been developed in ad hoc manner and a lot of work has focused on different aspects of crowdsourcing, such as computational techniques and performance analysis. The literature on crowdsourcing can be categorized into application, algorithm, performance and dataset. Fig. 1 shows a taxonomy of crowdsourcing.

The rest of this paper is organized as follows. Section II presents the study on crowdsourcing applications. It describes the categories and the characteristics of crowdsourcing applications. It presents how the applications in the previous works be categorized based on our categorization. Section III examines the algorithms developed for crowdsourcing systems. Section IV presents the survey on the performance aspects on evaluating the crowdsouring systems. Section V describes the experimental datasets available on the Web. Section VI gives a discussion and conclusion of our work.

II. Application

Because of the popularity of Web 2.0 technology, crowdsourcing websites attract much attentions at present [91], [90]. A crowdsourcing site has two groups of users: requesters and workers. The crowdsourcing site exhibits a list of available tasks, associating with reward and time period, that are presented by requesters; and during the period, workers compete to provide the best submission. Meanwhile, a worker selects a task from the task list and completes the task because the worker wants to earn the associated reward. At the end of the period, a subset of submissions are selected, and the corresponding workers are granted the reward by the requesters. In addition to monetary reward, a worker gains credibility when his task accepted by the requester. Sometimes, the task requester is obligated to pay every worker who has fulfilled the task according to the requirements. In some cases, workers are not motivated by rewards, but they work for fun or altruis [66]. In this section, we group crowdsourcing applications into four categories, and they are voting system, information sharing system, game and creative system.

A. Voting System

An example of popular crowdsourcing websites are Amazon Mechanical Turk (or MTurk) [1]. A large number of applications or experiments were conducted in Amazon's MTurk site. It can support a large number of voting tasks. These voting tasks require a crowdsourcing worker to select his answer from a number of choices. The answer that the majority selected is considered to be correct. Voting can be used as a tool to evaluate the correctness of an answer from the crowd. Some examples are shown below:

- **Geometric reasoning tasks** - The ability to interpret and reason about shapes is a specific human capability that has proven difficult to reproduce algorithmically. Some

978-0-7695-4578-3/11 $26.00 © 2011 IEEE
DOI

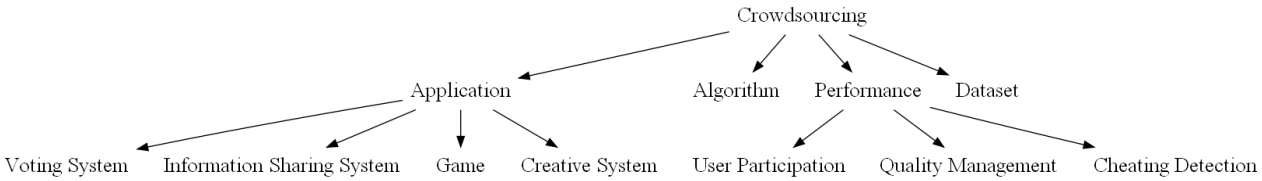

Fig. 1. A taxonomy of crowdsourcing

work were proposed to solve the problem of geometric reasoning on MTurk [39], [28].

- **Named entity annotation** - Named entity recognition is used to identify and categorize textual references to objects in the world, such as persons and organizations. MTurk is a very promising tool for annotating large-scale corpora, such as Arabic nicknames, Twitter data, large email datasets and medical named entities [29], [20], [49], [89].
- **Opinions** - Opinions are subjective preferences. Gathering opinions from the crowd can be achieved easily in a crowdsourcing system. Mellebeek et al. [59] used the crowdsourcing paradigm to classify Spanish consumer comments. They demonstrated that non-expert MTurk annotations outperformed expert annotations using a varity of classifiers.
- **Commonsense** - Obviously, humans can poss commonsense knowledge about the world, but computer programs cannot. Many studies focused on collecting commonsense knowledge in MTurk [23], [84].
- **Relevance evaluation** - Humans have to read through every document in a corpus to determine its relevance to a set of test queries. Alonso et al. [6] proposed crowdsourcing for relevance evaluation, so that each crowdsourcing work perform a small evaluation task.
- **Natural language annotation** - Natural language annotation is a task that is easy for humans but currently difficult for automated processes. Recently, researchers investigated MTurk as a source of non-expert natural language annotation, which is a cheap and quick alternative to expert annotations [4], [11], [21], [41], [65], [72]. Akkaya et al. [4] showed that crowdsourcing for subjectivity word sense annotation is reliable. Callison-Burch and Dredze [11] demonstrated their success on creating data for speech and language applications with a very low cost. Gao and Vogel [21] proved that crowdsourcing workers outperformed experts on word alignment tasks in terms of alignment error rate. Jha et al. [41] showed that it is possible to build up an accurate prepositional phrase attachment corpus by crowdsourcing workers. Parent and Eskenazi [65] demonstrated a way to cluster a task of dictionary definitions in MTurk. Skory and Eskenazi [72] submitted open cloze tasks to MTurk workers and discussed ways to evaluate the quality of the results of these tasks.
- **Spam identification** - Junk email cannot be determined without the task of understanding content by humans. Some anti-spam mechanisms such as Vipul's *Razor*[1] use human votes to determine if a given email is spam.

B. Information Sharing System

Websites can help to share information easily among Internet users. Some crowdsourcing systems aim to share various types of information among the crowd. For monitoring noise pollution, Maisonneuve [53] designed a system called NoiseTube which enables citizens to measure their personal exposure to noise in their everyday environment by using GPS-equipped mobile phones as noise sensors. The geo-localised measures and user-generated meta-data can be automatically sent and shared online with the public to contribute to the collective noise mapping of cities. Moreover, Choffnes et al. [15] utilized the crowdsoured contributions to monitor service-level network events and studied the impacts of network events on services in the view of end users. Furthermore, a lot of popular information sharing systems were launched on the Internet as shown in the following:

- *Wikipedia*[2] are online encyclopedias that are written by Internet users, and the writing is distributed in that essentially almost anyone can contribute to the Wiki.
- *Yahoo! Answers*[3] is a general question-answering forum to provide automated collection of human reviewed data at Internet-scale. These human-reviewed data are often required by enterprise and web data processing.
- *Yahoo! Suggestion Board*[4] is an Internet-scale feedback and suggestion system.
- The website *43Things*[5] also collects goals from users, and in turn it provides a way for users to find other users who have the same goals, even if they are uncommon.
- *Yahoo's flickr*[6] is a popular photo-sharing site and provides a mechanism for users to caption their photos. These captions are already being used as alternative text.
- *del.icio.us*[7] is a social bookmark site on the Internet developed by Golder and Huberman[22].

[1]Vipuls razor web site, http://sourceforge.net/projects/razor
[2]The free encyclopedia, http://en.wikipedia.org
[3]Yahoo! answers, http://answers.yahoo.com/
[4]Yahoo! Suggestion Board, http://suggestions.yahoo.com/
[5]43things website for collecting goals from users, http://www.43things.com/
[6]Yahoo's flickr, http://www.flickr.com/
[7]del.icio.us, http://del.icio.us/

C. Game

The concept of "Social Game" was pioneered by Luis Von Ahn and his colleagues, who created games with a purpose [76]. The games produce useful metadata as a by-product. By taking advantage of people's desire to be entertained, problems can be solved efficiently by online game players.

The online ESP Game [77] was the first human computation system, and it was subsequently adopted as the *Google Image Labeler*[8]. Its objective is to collect labels for images on Web. In addition to image annotation, the Peekaboom system [81] can help determine the location of objects in images, and the Squigl system [2] provides complete outlines of the objects in an image. Besides, Phetch [78], [79] provides image descriptions that improve web accessibility and image searches, while the Matchin system [2] helps image search engines rank images based on which ones are the most appealing. The concept of the ESP Game has been applied to other problems. For instance, the TagATune system [48], MajorMiner [54] and The Listen Game [75] provide annotation for sounds and music which can improve audio searches. The Verbosity system [80] and the Common Consensus system [51] collect commonsense knowledge that is valuable for commonsense reasoning and enhancing the design of interactive user interfaces. Green et al. [26] proposed PackPlay to mine semantic data. Examples in social annotation were described in [8], [9], [85]. Several GWAP-based geospatial tagging systems have been proposed in recent years, such as *MobiMission* [25], *Gopher* game [13] and *CityExplorer* [57], [58]. To simplify the way of designing a social game for a specific problem, Chan et al. [14] presented a formal framework for designing social games in general.

D. Creative System

The role of human in creativity cannot be replaced by any advanced technologies. The creative tasks, such as drawing and coding, can only be done by humans. As a result, some researchers seeked for crowdsourcing workers to do some creative tasks to reduce the production costs. An example is the Sheep Market. The Sheep Market is a web-based artwork to implicate thousands of online workers in the creation of a massive database of drawings. It is a collection of 10,000 sheeps created by MTurk workers, and each worker was paid US$0.02 to draw a sheep facing left [3], [46]. Another example is Threadless[9]. Threadless is a platform of collecting graphic t-shirt designs created by the communicty. Although technological advances rapidly nowadays, humans can innovate creative ideas in a product design process but computers cannot. It has no clue about how to solve a specific problem for developing a new product. Different individuals may create different ideas such as designing a T-shirt [10]. Moreover, Leimeister et al. [50] proposed to crowdsource software development tasks as ideas competitions to motivate more users to support and participate. Nowadays, scientists are being confronted by increasingly complex problems, but current technology unable to provide solutions. Some crowdsourcing systems were designed to solve these problems. Foldit[10] is a revolutionary new computer game that allows players to assist in predicting protein structures, an important area of biochemistry that seeks to find cures for diseases, by taking advantage of humans' puzzle-solving intuitions.

A crowdsourcing system typically supports only simple, independent tasks, such as labeling an image or judging the relevance of a search result. Some works proposed an idea of coordination among many individuals to complete more complex human computation tasks [52], [44]. Little et al. presented TurKit [52], which is a toolkit for exploring human computation algorithms on MTurk. TurKit allows users to write algorithms in a straight-forward imperative programming style, abstracting MTurk as a function call. Rather than solving many small, unrelated tasks partitioned into individual HITs, TurKit presented the notion that a single task, such as sorting or editing text, might require multiple coordinated HITs, and offers a persistence layer that makes it simple to iteratively develop such tasks without incurring excessive HIT costs. Kittur et al. presented CrowdForge [44] a general purpose framework for micro-task markets that provides a scaffolding for more complex human computation tasks which require coordination among many individuals, such as writing an article. CrowdForge abstracts away many of the programming details of creating and managing subtasks by treating partition/map/reduce steps as the basic building blocks for distributed process flows, enabling complex tasks to be broken up systematically and dynamically into sequential and parallelizable subtasks.

III. Algorithm

An algorithm can help to formalize the design of a crowdsourcing system. Yan et al. [87] designed an accurate real-time image search system for iPhone called CrowdSearch. CrowdSearch combines automated image search with real-time human validation of search results using MTurk. Nevertheless, an algorithm can model the performance of a crowdsourcing system. Wang et al. [83] modeled the completion time as a stochastic process and build a statistical method for predicting the expected time for task completion on MTurk. The experimental results showed that how time-independent variables of posted tasks (e.g., type of the task, price of the HIT, day posted, etc) affect completion time. Singh [71] proposed a game-theoretic framework for studying user behavior and motivation patterns in social media networks. DiPalantino and Vojnovic [16] modeled a competitive crowdsourcing system, and evidenced that participation rates are logarithmically increasing as a function of the offered reward. Ipeirotis et al. [38] presented an algorithm for quality management of the labeling process in a crowdsourcing system. The algorithm can generate a scalar score representing the inherent quality of each worker. Carterette and Soboroff [12] presented eight

[8]Google Image Labeler, http://images.google.com/imagelabeler/

[9]Threadless website: http://www.threadless.com/

[10]Foldit website: http://www.fold.it

models of possible errors for relevance judgments from crowd and showed how each affects an estimate of average precision. Jain and Parkes [40] surveyed existing game-theoretic models for various human computation designs, and also outlined the research challenges by advancing the game theory to enable better design of human computation systems.

IV. Performance

In addition to designing new applications and algorithms for the concept of crowdsourcing, several studies have investigated the performance aspect of crowdsourcing recently. These works can be categorized into user participation, quality management and cheating detection as shown in this section.

A. User Participation

In a crowdsourcing system, tasks are distributed to a population of anonymous Internet users for completion. Understanding the demographics of crowdsourcing workers and examining their behavior attracted significant attentions.

1) Demographics: As Amazon's Mechanical Turk (MTurk) is increasingly popular recently, Ross et al. [68] described how the worker population has changed over time, shifting from a primarily moderate-income, U.S.-based workforce towards an increasingly international group with a significant population of young, well-educated Indian workers. For these Indian workers, MTurk may increasingly function as a part- or full-time job. To disqualify the workers who participate but do not take the study tasks seriously, Downs et al. [17] screened MTurk workers by using two previously pilot tested screening questions. Experimental results showed that those that are professionals, students, and non-workers seem to be more likely to take the task seriously than financial workers, hourly workers, and other workers. Besides, men over 30 and women of any age were much more likely to qualify.

2) Financial Incentives: Numerous studies investigated the motivation of workers in crowdsourcing systems. Silberman et al. [70] presented that the importance of money compared to other motivations, with most respondents reporting they do not do tasks for fun or to kill time. 25 percent of Indian respondents and 13 percent of U.S. respondents reported that MTurk is their primary source of income. The impact of the financial incentives on specified crowdsourcing tasks were studied [27], [42], [60], [31]. Harris [27] found that financial incentives actually encourage quality if the task is designed appropriately in resume review. Kazai [42] found that low pay conditions result in increased levels of unusable and spam labels for a large corpus of digitized books. Moreno et al. [60] concluded that the question-answering sites should function better (faster answers by filling faster the FAQ lists) with both long-term and short-term rewards. To balance between a companys production cost and its workers' reservation wage, Horton and Chilton [31] presented a labor supply model to estimate a worker's reservation wage.

3) Intrinsic Incentives: Although monetary crowdsourcing incentive is dominant, some crowdsourcing systems do not offer monetary rewards to their workers, such as YouTube. What are the motivations of contribution in these systems? We describe various incentives other than financial incentives in the following. In the case of YouTube, attention, measured by the number of downloads, is an important driver of contributions [36]. Obviously, there exists a correlation between the rate at which content is generated and the number of downloads in YouTube. A lack of attention leads to a decrease in the number of videos uploaded and the consequent drop in productivity, which in many cases asymptotes to no uploads whatsoever. Wu et al. [86] demonstrated that contributors in YouTube who stop receiving attention tend to stop contributing, while prolific contributors attract an ever increasing number of followers and their attention in a feedback loop. In question-answering sites, Nam et al. [61] found altruism, learning, and competency are frequent motivations for top answerers to participate, but that participation is often highly intermittent. Besides, they showed that higher levels of participation correlate with better performance. For the purpose of open bug reporting, Ko and Chilana [45] found that in the case of Mozilla, what Mozilla gained was a small pool of talented developers and a number of critical fixes before the release of Firefox 1.0. These power contributors that contribute to open source projects have no intention of becoming regular contributors; they just want a bug fixed or a feature implemented. For the purpose of program coding, Archak [7] presented an empirical analysis of determinants of individual performance in multiple simultaneous crowdsourcing contests using a unique dataset for the world's largest competitive software development community (TopCoder.com). It studied the effects of the reputation system currently used by TopCoder.com on behavior of contestants. From observation, high rated contestants face tougher competition from their opponents in the competition phase of the contest. In an online photo-sharing community, Nov et al. [62] showed that tenure in the community does affect participation, but that this effect depends on the type of participation activity.

4) Worker Behavior: Many previous works showed that user interfaces can affect the behavior of crowdsourcing workers. By analyzing the waiting time for the posted tasks on MTurk, Ipeirotis [37] found that workers are limited by the current user interface and complete tasks by picking the tasks available through one of the existing sorting criteria. In addition to user interfaces, other factors affecting the behavior of crowdsourcing workers were found in the literature. Grady and Lease [24] investigated human factors involved in designing effective tasks on MTurk for document relevant assessment. They found that many of the same workers completed tasks in multiple batches, compromising the experimental control and likely introducing effects of training or fatigue. However, MTurk cannot prevent this happens. It is necessary to ensure each experiment involves a different set of workers in order to increase the output accuracy. Besides MTurk, other crowdsourcing websites were studied in literature. In 2008, Yang et al. [88] observed several characteristics in workers' activity over time on one of the biggest crowdsourcing websites in China, Taskcn.com. It found that most workers become inac-

tive after only a few submissions, while others keep attempting tasks. They tend to select tasks where they are competing against fewer opponents to increase their chances of winning; or they tend to select tasks with higher expected rewards. Instead of public crowdsourcing, a firm can outsource tasks to its employees rather than assign tasks to specified employees. Based on quantifiable effort-level metrics, Stewart et al. [74] proposed a SCOUT ((S)uper Contributor, (C)ontributor, and (OUT)lier) model for describing user participation inside the enterprise (within a company's firewall) and showed that it is possible to achieve a more equitable distribution of 33-66-1.

B. Quality Management

In a crowdsourcing system, a requester has to decide how to break down a task into several small tasks. A central challenge in crowdsourcing systems is how should a task be designed so as to induce good output from workers. Several studies performed comprehensive experiments using real datasets to study the impacts of user behavior on the quality of human-reviewed data. Mason and Watts [56] showed that increased financial incentives increase the quantity, but not the quality, of work performed by crowdsourcing workers. It is necessary to derive a set of design principles for tasks on crowdsourcing systems to guarantee the output quality of workers.

1) Image Annotation: Using Amazon Mechanical Turk [1] as an example, Snow et al. compared the quality of non-expert annotations and existing gold standard labels for natural language tasks provided by expert labelers [73]. The results demonstrated that it is required to collect an average of 4 non-expert labels per item in order to emulate expert-level label quality, and that the annotation quality can be improved significantly after applying bias correction techniques. Sheng et al. [69] proposed an analysis to model the data quality using repeated labeling with a cost. They found that, with repeated labeling, it is possible to improve the data quality at low cost, especially when labels are noisy. Moreover, when the cost of processing the unlabeled data is not free, repeated labeling is preferable in that it is effective and robust in providing labels of good quality. In 2010, Nowak and Rüger [63] conducted a study about inter-annotator agreement for multi-label image annotation. Although they did not answer the question how many annotation sets of non-experts are necessary to obtain comparable results to expert annotators, they evidenced that different annotators judge the same data and the inter-annotator agreement among different annotators can ensure the quality.

2) Text Annotation: Rashtchian et al. [67] found that the use of a qualification test provides the highest improvement of quality of linguistic data collected in MTurk. Hsueh et al. [34] considered the difficult problem of classifying sentiment in political blog snippets. They identified and confirmed the utility of the three selection criteria for high-quality annotations in MTurk: noise level, sentiment ambiguity, and lexical uncertainty. In fact, label quality is affected by cognitive awareness of human knowledge. Feng et al. [19] carried out experiments and showed that for the same task turkers answered questions quite differently if they were provided different knowledge in advance. Local search relevance is limited to topical relevance and geographical aboutness. Paiement et al. [64] used interannotator agreement as a quality measure for MTurk labels and discussed a simple approach to select only the most reliable labels. Wikipedia improves through the aggregation of many contributors' efforts. Kittur and Kraut [43] showed that adding more editors to an article improved article quality only when they used appropriate coordination techniques and was harmful when they did not. Implicit coordination through concentrating the work was more helpful when many editors contributed, but explicit coordination through communication was not.

3) General Tasks: Some work focused on the quality management of general tasks [35], [82]. Huang et al. [35] introduced a general approach for automatically designing tasks on MTurk. They constructed models for predicting the rate and quality of work. These models were trained on worker outputs over a set of designs, and were then used to optimize a task's design. They demonstrated that their models can accurately predict the quality of output per unit task and generate different designs depending on the quality metric. Voyer et al. [82] presented a two-phase, hybrid model for generating training data. They used named entity recognition as an example. In the first phase, a trained annotator labels all named entities in a text irrespective of type. In the second phase, naive crowdsourcing workers complete binary judgment tasks to indicate the type(s) of each entity. Decomposing the data generation task in this way results in a flexible, reusable corpus that accommodates changes to entity type taxonomies. In addition, it makes efficient use of precious trained annotator resources by leveraging highly available and cost effective crowdsourcing worker pools in a way that does not sacrifice quality.

C. Cheating Detection

Due to the anonymity of crowdsourcing workers, malicious workers often try to maximise their financial gains by producing generic answers rather than actually working on the task. Currently, cheat-detection techniques are either based on control questions which are evaluated automatically or rely on manual checking by the requester. Eickhoff and de Vries [18] inspected the commonly observed methods of malicious crowdsourcing workers, such as task-dependent evaluation, interface-dependent evaluation and audience-dependent evaluation. Based on experimental results, they concluded that malicious workers are less frequently encountered in novel tasks that involve a degree of creativity and abstraction, and prior crowd filtering can greatly reduce the number of malicious workers.

For the crowdsourcing systems that control questions are not applicable and manual re-checking is ineffective, Hirth et al. [30] presented two crowd-based approaches to detect cheating workers: a majority decision (MD) and an approach using a control group (CG) to re-checking the main task. For MD, the same task is given to several different workers and the results are compared. The result which most of the workers submitted is assumed to be correct. For CG, a single worker

works on a main task and a control group consisting of certain other workers re-checks the result, whether it is valid or not. Usually the main task is expensive, while the re-check task is cheaper. A task is considered to be valid, if the majority of the control group decides the task is correctly done. Experimental results showed that crowd-based cheat-detection mechanisms are cheap, reliable, easy to implement, and their applicability to different types of typical crowdsourcing tasks.

For some suitations, hiring experts for fraud detection is very expensive. Almendra and Schwabe [5] proposed the use of crowdsourcing to improve precision and recall of current fraud detection techniques for online auction sites. They showed that they could distinguish fraudsters from common sellers before negative feedback arrived and looking just at a snapshot of seller profiles.

V. Dataset

A number of crowdsourcing datasets are now available for further research. For example, von Ahn et al. contributed a list of 100,000 images with English labels from their ESP Game[11]. Law et al. released the research dataset for a human computation game called TagATune[12] [48]. Their website contains human annotations collected by the TagATune game, the corresponding sound clips from a web site for downloading songs called Magnatune[13], the source code of the scripts, and a detailed analysis of the track's structure and musical content. Recently, Chen et al. developed the ESP Lite game, which is similar to the ESP game introduced by von Ahn et al., and collected statistics for players playing the game[14]. Oversity Ltd. developed CiteULike[15], a free website, to help academics keep track of the articles they are reading on. Users are encouraged to make their libraries publicly available on the web so others can get the benefit of discovering useful articles they might not otherwise have found. In 2009, Benjamin Markines and Filippo Menczer extracted relationships among tags and resources from the available datasets of two social bookmarking systems, Bibsonomy[16] and GiveALink[17] [55]. In 2010, Ipeirotis et al. gathered all avaliable information from Amazon Mechanical Turk by computing daily statistics for new projects and completed tasks once a day and shared the dataset to the public[18]. Körner and Strohmaier [47] posted a list of social tagging datasets made available for research[19].

[11] ESP Game dataset: http://server251.theory.cs.cmu.edu/ESPGame100k.tar.gz

[12] Tagatune Dataset website: http://tagatune.org/Magnatagatune.html

[13] Magnatune Website - Pay for downloading songs: http://magnatune.com

[14] The website of IIS-NRL Games With A Purpose - ESP Lite: http://hcomp.iis.sinica.edu.tw/dataset/dataset_esplite20100101.php

[15] CiteULike website: http://www.citeulike.org and the dataset website: http://svn.citeulike.org/svn/plugins/HOWTO.txt

[16] The Bibsonomy.org website: http://Bibsonomy.org and the dataset website: http://bibsonomy.org/help/doc/api.html

[17] The GiveALink.org website: http://GiveALink.org and the dataset website: http://givealink.org/main/download

[18] Mechanical Turk Tracker website: http://www.mturk-tracker.com/

[19] A List of Social Tagging Datasets Made Available for Research: http://kmi.tugraz.at/staff/markus/datasets/

VI. Conclusion

We have surveyed various crowdsourcing systems, and categorized them into four types: application, algorithm, performance and dataset. To the best of our knowledge, this is the first extensive survey of the emerging crowdsourcing issue. This survey not only provides a better understanding about crowdsourcing systems, but also facilitates future research activities and application developments in the field of crowdsourcing.

Acknowledgment

The authors would like to thank the reviewers for their helpful comments. This work was partially supported by a grant from Microsoft (Project No. CUHK 6902498) and the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK 413210). It is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

References

[1] Amazon mechanical turk. https://www.mturk.com/.

[2] GWAP. http://www.gwap.com/gwap/.

[3] The sheep market. http://www.thesheepmarket.com/.

[4] C. Akkaya, A. Conrad, J. Wiebe, and R. Mihalcea. Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 195–203, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[5] V. Almendra and D. Schwabe. Fraud detection by human agents: A pilot study. In *Proceedings of the 10th International Conference on E-Commerce and Web Technologies*, EC-Web 2009, pages 300–311, Berlin, Heidelberg, 2009. Springer-Verlag.

[6] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42:9–15, November 2008.

[7] N. Archak. Money, glory and cheap talk: analyzing strategic behavior of contestants in simultaneous crowdsourcing contests on topcoder.com. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 21–30, New York, NY, USA, 2010. ACM.

[8] M. Bernstein, D. Tan, G. Smith, M. Czerwinski, and E. Horvitz. Collabio: a game for annotating people within social networks. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, UIST '09, pages 97–100, New York, NY, USA, 2009. ACM.

[9] M. S. Bernstein, D. Tan, G. Smith, M. Czerwinski, and E. Horvitz. Personalization via friendsourcing. *ACM Trans. Comput.-Hum. Interact.*, 17:6:1–6:28, May 2008.

[10] D. C. Brabham. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence: The International Journal of Research into New Media Technologies*, 14(1):75–90, 2008.

[11] C. Callison-Burch and M. Dredze. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 1–12, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[12] B. Carterette and I. Soboroff. The effect of assessor error on ir system evaluation. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 539–546, New York, NY, USA, 2010. ACM.

[13] S. Casey, B. Kirman, and D. Rowland. The gopher game: a social, mobile, locative game with user generated content and peer review. In *International Conference on Advances in Computer Entertainment Technology*, pages 9–16, 2007.

[14] K. T. Chan, I. King, and M.-C. Yuen. Mathematical modeling of social games. In *CSE '09: Proceedings of the 2009 International Conference on Computational Science and Engineering*, pages 1205–1210. IEEE Computer Society, 2009.

[15] D. R. Choffnes, F. E. Bustamante, and Z. Ge. Crowdsourcing service-level network event monitoring. *SIGCOMM Comput. Commun. Rev.*, 40:387–398, August 2010.

[16] D. DiPalantino and M. Vojnovic. Crowdsourcing and all-pay auctions. In *Proceedings of the 10th ACM conference on Electronic commerce*, EC '09, pages 119–128, New York, NY, USA, 2009. ACM.

[17] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor. Are your participants gaming the system?: screening mechanical turk workers. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 2399–2402, New York, NY, USA, 2010. ACM.

[18] C. Eickhoff and A. P. de Vries. How crowdsourcable is your task? In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM 2011)*, WSDM 2011, pages 11–14, New York, NY, USA, 2011. ACM.

[19] D. Feng, S. Besana, K. Boydston, and G. Christian. Towards high-quality data extraction via crowdsourcing. In *Proceedings of the CrowdConf 2010*, CrowdConf 2010, 2010.

[20] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 80–88, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[21] Q. Gao and S. Vogel. Consensus versus expertise: a case study of word alignment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 30–34, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[22] S. A. Golder and B. A. Huberman. The structure of collaborative tagging systems. *CoRR*, abs/cs/0508082, 2005.

[23] J. Gordon, V. D. Benjamin, and L. K. Schubert. Evaluation of commonsense knowledge with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 159–162, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[24] C. Grady and M. Lease. Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 172–179, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[25] L. Grant, H. Daanen, S. Benford, A. Hampshire, A. Drozd, and C. Greenhalgh. MobiMissions: the game of missions for mobile phones. In *ACM SIGGRAPH*, 2007.

[26] N. Green, P. Breimyer, V. Kumar, and N. F. Samatova. Packplay: mining semantic data in collaborative games. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 227–234, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[27] C. G. Harris. You're hired! an examination of crowdsourcing incentive models in human resource tasks. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM 2011)*, WSDM 2011, pages 15–18, New York, NY, USA, 2011. ACM.

[28] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 203–212, New York, NY, USA, 2010. ACM.

[29] C. Higgins, E. McGrath, and L. Moretto. Mturk crowdsourcing: a viable method for rapid discovery of arabic nicknames? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 89–92, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[30] M. Hirth, T. Hoßfeld, and P. Tran-Gia. Cheat-detection mechanisms for crowdsourcing. Technical Report 474, University of Würzburg, 8 2010.

[31] J. J. Horton and L. B. Chilton. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce*, EC '10, pages 209–218, New York, NY, USA, 2010. ACM.

[32] J. Howe. The rise of crowdsourcing. *Wired*, 14(6), June 2006.

[33] J. Howe. *Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business*. Crown Business, 2008.

[34] P.-Y. Hsueh, P. Melville, and V. Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, HLT '09, pages 27–35, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[35] E. Huang, H. Zhang, D. C. Parkes, K. Z. Gajos, and Y. Chen. Toward automatic task design: a progress report. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 77–85, New York, NY, USA, 2010. ACM.

[36] B. A. Huberman, D. M. Romero, and F. Wu. Crowdsourcing, attention and productivity. *J. Inf. Sci.*, 35:758–765, December 2009.

[37] P. G. Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS*, 17:16–21, December 2010.

[38] P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 64–67, New York, NY, USA, 2010. ACM.

[39] A. P. Jagadeesan, A. Lynn, J. R. Corney, X. T. Yan, J. Wenzel, A. Sherlock, and W. Regli. Geometric reasoning via internet crowdsourcing. In *2009 SIAM/ACM Joint Conference on Geometric and Physical Modeling*, SPM '09, pages 313–318, New York, NY, USA, 2009. ACM.

[40] S. Jain and D. C. Parkes. The role of game theory in human computation systems. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '09, pages 58–61, New York, NY, USA, 2009. ACM.

[41] M. Jha, J. Andreas, K. Thadani, S. Rosenthal, and K. McKeown. Corpus creation for new genres: A crowdsourced approach to pp attachment. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 13–20, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[42] G. Kazai. An exploration of the influence that task parameters have on the performance of crowds. In *Proceedings of the CrowdConf 2010*, CrowdConf 2010, 2010.

[43] A. Kittur and R. E. Kraut. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, CSCW '08, pages 37–46, New York, NY, USA, 2008. ACM.

[44] A. Kittur, B. Smus, and R. Kraut. Crowdforge: crowdsourcing complex work. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, CHI EA '11, pages 1801–1806, New York, NY, USA, 2011. ACM.

[45] A. J. Ko and P. K. Chilana. How power users help and hinder open bug reporting. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 1665–1674, New York, NY, USA, 2010. ACM.

[46] A. M. Koblin. The sheep market. In *Proceeding of the seventh ACM conference on Creativity and cognition*, C&C '09, pages 451–452, New York, NY, USA, 2009. ACM.

[47] C. Körner and M. Strohmaier. A call for social tagging datasets. *SIGWEB Newsl.*, pages 2:1–2:6, January 2010.

[48] E. Law and L. von Ahn. Input-agreement: A new mechanism for data collection using human computation games. In *ACM CHI*, 2009.

[49] N. Lawson, K. Eustice, M. Perkowitz, and M. Yetisgen-Yildiz. Annotating large email datasets for named entity recognition with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 71–79, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[50] J. Leimeister, M. Huber, U. Bretschneider, and H. Krcmar. Leveraging crowdsourcing: Activation-supporting components for it-based ideas competition. *J. Manage. Inf. Syst.*, 26:197–224, July 2009.

[51] H. Lieberman, D. Smith, and A. Teeters. Common Consensus: a web-based game for collecting commonsense goals. In *ACM Workshop on Common Sense for Intelligent Interfaces*, 2007.

[52] G. Little, L. B. Chilton, M. Goldman, and R. C. Miller. Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pages 57–66, New York, NY, USA, 2010. ACM.

[53] N. Maisonneuve, M. Stevens, M. E. Niessen, P. Hanappe, and L. Steels. Citizen noise pollution monitoring. In *Proceedings of the 10th Annual International Conference on Digital Government Research: Social Networks: Making Connections between Citizens, Data and Government*, dg.o '09, pages 96–103. Digital Government Society of North America, 2009.

[54] M. Mandel and D. Ellis. A web-based game for collecting music metadata. In *8th International Conference on Music Information Retrieval (ISMIR)*.

[55] B. Markines and F. Menczer. A scalable, collaborative similarity measure for social annotation systems. In *HT '09: Proceedings of the 20th ACM conference on Hypertext and hypermedia*, pages 347–348, New York, NY, USA, 2009. ACM.

[56] W. Mason and D. J. Watts. Financial incentives and the "performance of crowds". *SIGKDD Explor. Newsl.*, 11:100–108, May 2010.

[57] S. Matyas. Playful geospatial data acquisition by location-based gaming communities. *The International Journal of Virtual Reality*, 6(3):1–10, 2007.

[58] S. Matyas, C. Matyas, C. Schlieder, P. Kiefer, H. Mitarai, and M. Kamata. Designing Location-based Mobile Games With A Purpose: Collecting Geospatial Data with CityExplorer. In *ACM ACE*, 2008.

[59] B. Mellebeek, F. Benavent, J. Grivolla, J. Codina, M. R. Costa-jussà, and R. Banchs. Opinion mining of spanish customer comments with non-expert annotations on mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 114–121, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[60] A. Moreno, J. L. de la Rosa, B. K. Szymanski, and J. M. Barcenas. Reward system for completing faqs. In *Proceeding of the 2009 conference on Artificial Intelligence Research and Development: Proceedings of the 12th International Conference of the Catalan Association for Artificial Intelligence*, pages 361–370, Amsterdam, The Netherlands, The Netherlands, 2009. IOS Press.

[61] K. K. Nam, M. S. Ackerman, and L. A. Adamic. Questions in, knowledge in?: a study of naver's question answering community. In *Proceedings of the 27th international conference on Human factors in computing systems*, CHI '09, pages 779–788, New York, NY, USA, 2009. ACM.

[62] O. Nov, M. Naaman, and C. Ye. Analysis of participation in an online photo-sharing community: A multidimensional perspective. *J. Am. Soc. Inf. Sci. Technol.*, 61:555–566, March 2010.

[63] S. Nowak and S. Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, pages 557–566, New York, NY, USA, 2010. ACM.

[64] J.-F. Paiement, J. G. Dr. Shanahan, and R. Zajac. Crowdsourcing local search relevance. In *Proceedings of the CrowdConf 2010*, CrowdConf 2010, 2010.

[65] G. Parent and M. Eskenazi. Clustering dictionary definitions using amazon mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 21–29, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[66] A. J. Quinn and B. B. Bederson. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 1403–1412, New York, NY, USA, 2011. ACM.

[67] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 139–147, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[68] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, CHI EA '10, pages 2863–2872, New York, NY, USA, 2010. ACM.

[69] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *ACM KDD*, 2008.

[70] M. S. Silberman, L. Irani, and J. Ross. Ethics and tactics of professional crowdwork. *XRDS*, 17:39–43, December 2010.

[71] V. K. Singh, R. Jain, and M. S. Kankanhalli. Motivating contributors in social media networks. In *Proceedings of the first SIGMM workshop on Social media*, WSM '09, pages 11–18, New York, NY, USA, 2009. ACM.

[72] A. Skory and M. Eskenazi. Predicting cloze task quality for vocabulary training. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, IUNLPBEA '10, pages 49–56, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[73] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Morristown, NJ, USA, 2008. Association for Computational Linguistics.

[74] O. Stewart, D. Lubensky, and J. M. Huerta. Crowdsourcing participation inequality: a scout model for the enterprise domain. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 30–33, New York, NY, USA, 2010. ACM.

[75] D. Turnbull, R. Liu, L. Barrington, and G. Lanckriet. A game-based approach for collecting semantic annotations of music. In *8th International Conference on Music Information Retrieval (ISMIR)*.

[76] L. von Ahn. Games with a Purpose. *IEEE Computer*, 39(6):92–94, June 2006.

[77] L. von Ahn and L. Dabbish. Labeling Images with a Computer Game. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, 2004.

[78] L. von Ahn, S. Ginosar, M. Kedia, and M. Blum. Improving Image Search with PHETCH. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[79] L. von Ahn, S. Ginosar, M. Kedia, R. Liu, and M. Blum. Improving accessibility of the web with a computer game. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, 2006.

[80] L. von Ahn, M. Kedia, and M. Blum. Verbosity: A Game for Collecting Common-Sense Facts. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, 2006.

[81] L. von Ahn, R. Liu, and M. Blum. Peekaboom: A Game for Locating Objects in Images. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, 2006.

[82] R. Voyer, V. Nygaard, W. Fitzgerald, and H. Copperman. A hybrid model for annotating named entity training corpora. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 243–246, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[83] J. Wang, S. Faridani, and P. G. Ipeirotis. Estimating the completion time of crowdsourced tasks using survival analysis models. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM 2011)*, WSDM 2011, pages 31–34, New York, NY, USA, 2011. ACM.

[84] R. Wang and C. Callison-Burch. Cheap facts and counter-facts. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 163–167, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[85] L. Weng and F. Menczer. Givealink tagging game: an incentive for social annotation. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 26–29, New York, NY, USA, 2010. ACM.

[86] F. Wu, D. M. Wilkinson, and B. A. Huberman. Feedback loops of attention in peer production. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, pages 409–415, Washington, DC, USA, 2009. IEEE Computer Society.

[87] T. Yan, V. Kumar, and D. Ganesan. Crowdsearch: exploiting crowds for accurate real-time image search on mobile phones. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, MobiSys '10, pages 77–90, New York, NY, USA, 2010. ACM.

[88] J. Yang, L. A. Adamic, and M. S. Ackerman. Crowdsourcing and knowledge sharing: strategic user behavior on taskcn. In *Proceedings of the 9th ACM conference on Electronic commerce*, EC '08, pages 246–255, New York, NY, USA, 2008. ACM.

[89] M. Yetisgen-Yildiz, I. Solti, F. Xia, and S. R. Halgrim. Preliminary experience with amazon's mechanical turk for annotating medical named entities. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 180–183, Morristown, NJ, USA, 2010. Association for Computational Linguistics.

[90] M.-C. Yuen, L.-J. Chen, and I. King. A survey of human computation systems. In *CSE '09: Proceedings of IEEE International Conference on Computational Science and Engineering*, pages 723–728. IEEE Computer Society, 2009.

[91] M.-C. Yuen, I. King, and K.-S. Leung. Task matching in crowdsourcing. In *CPSCom '11: Proceedings of The 4th IEEE International Conferences on Cyber, Physical and Social Computing*. IEEE Computer Society, 2011. To be appeared.