Expert Systems with Applications 41 (2014) 7987–7994

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Brief survey of crowdsourcing for data mining

Guo Xintong [a], Wang Hongzhi [a,*], Yangqiu Song [b], Gao Hong [a]

[a] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150006, China
[b] Department of Computer Science and Technology, University of Illinois at Urbana-Champaign, IL, USA

ARTICLE INFO

Article history:
Available online 11 July 2014

Keywords:
Data mining
Crowdsourcing
Quality control
Survey

ABSTRACT

Crowdsourcing allows large-scale and flexible invocation of human input for data gathering and analysis, which introduces a new paradigm of data mining process. Traditional data mining methods often require the experts in analytic domains to annotate the data. However, it is expensive and usually takes a long time. Crowdsourcing enables the use of heterogeneous background knowledge from volunteers and distributes the annotation process to small portions of efforts from different contributions. This paper reviews the state-of-the-arts on the crowdsourcing for data mining in recent years. We first review the challenges and opportunities of data mining tasks using crowdsourcing, and summarize the framework of them. Then we highlight several exemplar works in each component of the framework, including question designing, data mining and quality control. Finally, we conclude the limitation of crowdsourcing for data mining and suggest related areas for future research.

1. Introduction

People from different fields analyze a variety of datasets to understand human behaviors, find new trends in society, and possibly formulate adequate policies in response. Typically, we address the problem of finding interesting and unknown patterns via data mining methodology. Data mining enables people to extract information from a data set and convert it into a comprehensible structure for further use.

Typical data mining techniques, however, are not suitable for current applications. First, when mining the datasets, we must have access to all relevant information. In fact, it is impossible to obtain all these transactions, which mainly because of the properties of the human memory. People's memories are prone to remember summaries, rather than exact details (Boim et al., 2012). Consider the following case. A social scientist wants to analyze life habits of people. The database includes leisure activities (watching TV, jogging, reading, etc.) correlated with time of the day, weather and so on. But it is unrealistic for people to recall an exhaustive list of all cases they did. People can make assumptions in order to compensate the loss of information by crowdsourcing the mining task. Second, some mining algorithms are time-consuming, especially used for large datasets, which also leads to much more extra cost. Finally, raw data mining technologies are lack of related information. Algorithm has to be taught the knowledge before mining. For example, for the classification problem, labeled data is used for training the classifier to have the ability of classifying new coming test data. However, acquiring the labeled data is time consuming and costly.

In the circumstances, we can solve this problem by crowdsourcing. As crowdsourcing is based on the people who have the incentives to work on small tasks, the mining tasks can benefit from the aggregation of labeling work which is time-controllable, flexible, easy to implement due to the current crowdsourcing platform.

Crowdsourcing is an emerging and powerful information procurement paradigm that has appeared under many names, including social computing, collective intelligence and human computation (Quinn & Bederson, 2011). Requesters decompose the whole task into several small tasks and push them to the crowd, and workers accomplish questions for intrinsic or extrinsic reasons (Von Ahn & Dabbish, 2008). Although people may not remember all of transactions precisely, many current studies prove that simple summaries can still achieve a positive result, and even more complicated questions (Boim et al., 2012).

Crowdsourcing has played important roles in data mining. In some kind of scenarios, it can help people resolve the problems in a more efficient way and give them deeply understanding to apply crowdsourcing. Here we give some situations for the applications of crowdsourcing techniques in various real-world data mining tasks.

Crisis Map: Crisis map is one of the most representative applications of crowdsourcing. It is a platform, designed to do information collection, analysis of mass data and display in a straightforward way in real time during a crisis. It has become a powerful mechanism for a large number of people to contribute about crisis events.

* Corresponding author.
E-mail address: wangzh@hit.edu.cn (W. Hongzhi).

http://dx.doi.org/10.1016/j.eswa.2014.06.044
0957-4174/

People not only provide useful information about the crisis situation, but also cluster materials into meaningful categories. Then people with no field-specific skills filter out the irrelevant parts, do analysis and assemble reports. What is more, these crisis maps can visualize a large amount of data and give the rescue teams better insights of the relief situation (Goolsby, 2010). People generated numerous messages and photos after the devastating earthquake in Haiti happened on 2010, through social media networking (Gao, Barbier, & Goolsby, 2011). The use of crisis map in disaster accelerates the application development to leverage the value of crowdsourcing.

Homeland Security: Crowdsourcing can also benefit homeland security. We can use crowds to contribute to delivering quality information and identifying the suspects. The Boston Marathon bombing, happened on April 15, 2013, caused many injuries and deaths. On the next day, an appeal went out to the public, urging the citizens to submit all photos and videos that they might have of the Boston Marathon environment (Spenser, 2013). A number of sites (Reddit & 4chan) were set up to aggregate the photos and videos. And then the crowd helped to identify the suspects in the flooded materials collected from the first appeal. The public responded quickly and provided valuable intelligence both times (Markowsky, 2013).

Facebook: Facebook is another popular website that can be used for crowdsourcing. Compared to twitter, Facebook has more information sources, including blog and picture. So Facebook can fulfill some sophisticated tasks, such as character analysis, financial analysis (Libert & Spector, 2007), activity planning (Brabham, Sanchez, & Bartholomew, 2009), and product repository generation (Budde & Michahelles, 2010). Facebook builds up various applications for individuals to design their own crowdsourcing tasks.

Lots of crowdsourcing platforms have sprung up during these years, such as CloudCrowd (used to write and edit the project), Crowdflower and so on. The Amazon Mechanical Turk is one of the most famous and largest in scale. The Amazon Mechanical Turk (MTurk) allows individuals or business corporations (known as requester) post various tasks, such as image clustering, document labeling, creative designing and so on. The workers (known as Turker) choose HITs (Human Intelligence Tasks) to accomplish for monetary incentive. The requesters connect web applications and MTurk through open interfaces (APIs), which benefit customizing task design and analysis.

Managing and analysis data on the crowdsourcing platform have recently become a wide-spread phenomenon, leading to explosion of research activity in recent years (Amsterdamer, Grossman, Milo, & Senellart, 2013a, 2013b). What we cannot ignore are the challenges arising from the real-world applications. The challenges include, but not limit to adaptive question deliver system, recommendation framework for requester to design task, as well as specific mining algorithm. We will discuss them in detail in Section 7.

Apparently, a better understanding in crowdsourcing for data mining can help us tap into this powerful new resource in a more efficient way. That is why our survey is important. The contributions of this paper are summarized as follows.

1. We review the state-of-the-art work on the crowdsourcing for data mining in recent years. As we know, this is the first survey about crowdsourcing techniques for data mining.
2. We point it out the difference between raw data mining algorithms and those based on crowdsourcing. There are more factors to be considered when data mining task accomplished by crowdsourcing.
3. According to existing work, we summarize a general framework of crowdsourcing for data mining, which includes question design, mining process and quality control.
4. We review the highlight works in each component of the framework.
5. We give some generic tips about task design and discuss the quality control method selection strategies for data mining tasks. It is quite instructive and meaningful for requester to follow.
6. We investigate challenges and opportunities of data mining tasks in crowdsourcing. We also conclude the limitation of crowdsourcing for data mining and suggest related areas for future research.

Crowdsourcing is a powerful tool for government to collect and analyze data. It provides new opportunities for data mining which has widely applications in expert systems. The methods proposed in this article are not only fit for data mining, but also good references for other related work, such as information retrieval, machine learning and crisis management. Such fields also have close relationship to expert systems.

The rest of this paper is organized as follows. Section 1 provides a general framework for crowd mining. Section 2 tells how to design a task for data mining. Section 3 proposes various data mining tasks that can be performed by means of crowdsourcing. Section 4 presents the study on quality control for data mining, which is closely related to the result of data mining. Section 5 introduces situations that crowdsourcing method is not suitable for tackling tasks. Section 6 describes what we can do in the future. At the end of the article, we give a discussion and conclusion of our work.

2. Framework

Traditional data mining methodologies and technologies are sometimes time-consuming, inflexible, expensive to implement, and poor scalable. Crowdsourcing can be applied to manage data and extract interesting patterns from the data sets more efficiently and intelligently by comparison. From existing work of crowdsourcing techniques (Boim et al., 2012; Amsterdamer et al., 2013a, 2013b; Barbier, Zafarani, Gao, Fung, & Liu, 2012; Karger, Oh & Shah, 2011; Weaver, Boyle & Besaleva, 2012), we conclude that using crowdsourcing for data mining can be performed by following a three-step procedure: question design, mining and quality control.

Question Design: Well-designed tasks can obtain high-quality answers. Questions should be designed based on the purpose of the data mining task. We address the problem of effective crowdsourcing, namely gathering data from the crowd in a way that is economical in time and expense.

Mining: The mining phase absolutely takes the center stage in the whole process. Data mining tasks can be divided into the multiple kinds: classification, clustering, semi-supervised learning, and association rules mining. Classification has been widely used in many fields, such as face recognition, disaster rescue. Some research takes advantages of crowdsourcing to identify association rules between relating signs and symptoms to diseases (Wright, Chen, & Maloney, 2010). Crowdsourcing appears to have several important merits compared with other automatic knowledge-based approaches.

Quality Control: Due to the nature of crowdsourcing task, a quality control step is necessary for the result after the mining step. Malicious workers, who are only attempting to maximize their income or lack of necessary training, are detrimental to the mining result (Venetis & Garcia-Molina, 2012). Quality control step uses vote system, redundant workers, worker's reputation and other methods to pick out the irresponsible workers.

In the following sections, we will discuss these three steps, respectively.

3. Question design

In a crowdsourcing system, a requester has to decide how to split the whole task into several small tasks, each of which could be distributed as a unit. For some kind of data mining tasks, such as clustering, each worker has only a partial view of the data. How to decompose the whole task into small pieces is a primary factor to be considered.

According to the data mining task, we should ask the correct questions to achieve good performance, and design different questions to different workers to make further progress. In this process, we address the problem of gathering data from the crowd with fewer questions and more information.

Here Alonso and Lease (2011) give some generic tips for task design.

1. Experiments should be self-contained.
2. Instructions for task must be short and simple, brief and concise. Too complex task or ambiguous instructions for workers to understand will generate poor quality answers.
3. Be very clear with the relevance task. If a similar task has been published, there is no need to replicate it.
4. Engage with the worker. Cancel useless stuff. We can run a test with a very small data set and gather feedback from workers to enhance the task design.
5. Always ask for feedback (open-ended question) in an input box. Iterate and modify accordingly.
6. UI design. Highlight important concepts and pay attention to the layouts.

Eickhoff and de Vries (2011) investigate the commonly cheating methods of malicious crowdsourcing workers. They establish a series of experimental schemes, including task-dependent valuation, interface-dependent evaluation and audience-dependent evaluation. They draw a conclusion that bad workers are rarely appeared in novel tasks that contain innovation and information extraction, and the number of poor-quality workers will be considerably reduced if we apply filtering process.

Besides, the properties of the question are closely related to the workers' answers. Well designed task is conductive to high quality mining result. Defensive task design is a tool for quality control, such as employing spammers to judge the result and referring to worker's reputation. We can add qualifying questions which can block the unqualified workers, gold standard questions for which malicious answers can be filtered out in the process of crowdsourcing, and checking completion time. The easiest implementation method to distinguish correct answer is to hire redundant workers to finish the same HIT, and then aggregate them by applying a majority rule (Kazai, Kamps, Koolen, & Milic-Frayling, 2011).

4. Mining data from crowdsourcing

Various types of data mining tasks can be accomplished by means of crowdsourcing, e.g. classification, clustering, semi-supervised learning, and association rule mining. Traditional algorithms have difficulties in tackling these problems, for the lack of knowledge. In these situations, the powerful crowds can perform more accurately, flexibly and efficiently than the existing automatic algorithms. We will discuss how crowdsourcing can be used to solve these problems and provide some application scenario.

4.1. Classification

Crowdsourcing can be utilized to solve classification problem. Crowdsourcing method has much more advantages over the general data mining technology. A typical classification problem is to distinguish male and female from a social network users. In the original data mining perspective, a classifier is built to extract features from the given datasets in the first step. In the second step, the classifier is used for classification. Compared to this, if we distribute this task to the crowd, everyone can classify the objects immediately. Sometimes we can say that the wisdom of the crowd allows for more accuracy than any other classification algorithm.

Documents categorization can be accomplished by users on the website as well. This approach has been applied on many domains successfully, such as Digg and Yahoo! Directory. Digg is an aggregator to customize the user's news front page. Digg also allows users to tag to the submitted links, which widen the scope to include more relevant articles the users may be interested in. We find that accuracy increases as the more and more users participate in.

Another widely known example is CAPTCHA. *"CAPTCHA* (Von Ahn, Maurer, McMillen, Abraham, & Blum, 2008) *is a challenge-response test used on the World Wide Web to determine whether a user is a human or a computer".* Technologies at present cannot recognize distorted text as fast as humans can. Human enter the characters to digitize handwriting text (Von Ahn, 2009).

Chilton, Little, Edge, Weld, and Landay (2013)) present an algorithm, named CASCADE, to create an overall consistent taxonomy by distributing HITs to many individuals, each of whom has only a partial view of the data. CASCADE hires many unskilled labors to produce taxonomies. The quality of classification is approximate to that of human experts, while the cost of CASCADE is very cheap. Furthermore, Bragg and Weld (2013) present DELUGE, an improved workflow on the basis of CASCADE. DELUGE produces taxonomies with equivalent quality in spite of reducing the workforce. The categorization step, which is the most consuming, is optimized by decision theory. The experiment result demonstrates that less than 10% of the workers are required by the original approach.

As we mentioned before, crowdsourcing helps a lot in disaster rescue. Zhai, Kijewski-Correa, Hachen, and Madey (2012) establish an online framework that generates human computation resources to tackle an image labeling task, classifying post-disaster photos according to damage extent. In real life, such type of information is needed to manage risks in disaster-prone areas, both in pre-disaster risk reductions and post-disaster damage assessments.

Other related works that considering the budget allocation contributed to the classification work as well. Tran-Thanh, Venanzi, Rogers, and Jennings (2013) raise the issue of how to allocate the budget for redundant workers when dealing with classification tasks, where the key challenge is to find a proper balance between the total cost and quality. They propose CrowdBudget, a budget allocation algorithm, aiming to minimize estimation error with the limited fund.

4.2. Clustering

Clustering is more complicated than classification problem. One of the aspects, there are lots of ways to define the similarity between items. Different measure of similarity may lead to different result. Similarly, we can perform cluster task on the crowdsourcing platform.

Many recent social networking sites give humans permission to create categories. In the case of Twitter, users assign tags to their tweets in order to follow up the trending topics. This facilitates quick retrieval when searching for tweets and again, and forms a discussion groups about the hot issue automatically (Barbier et al., 2012). What is more, a large set of Tweets relevant to a particular cluster can be an excellent source for professionals to analyze.

Chen, Wang, and Tan (2012) create a friendly environment, using crowd to visualize web images into clusters. The method has two stages. The first stage separates an image set into multiple clusters and the second stage purifies each generated cluster independently. During the whole stage, computers select informative images and the crowds help to label the images to improve the quality. The experimental results demonstrate the combinations of computers and a large number of human workers benefit high-quality visual clusters.

Here we require addressing some challenges in crowdsourced clustering. (1) Each worker has only a fraction of the data, so we need additional algorithm to merge the results. (2) Different workers may have different clustering standard, leading to produce different numbers of categories. (3) The underlying category structure may be hierarchical. According to the intractable problems mentioned above, Gomes, Welinder, Krause, and Perona (2011) propose a model, based on Variational Bayes method, of how crowdsourcing can be applied to clustering. First, divide the dataset into overlap subsets. And then workers propose partial clustering. Finally, use Bayes model to aggregate the partial clusters into one cluster.

4.3. Semi-supervised learning

In semi-supervised learning, we use labeled data to acquire necessary knowledge, and then label the unlabeled data. This procedure significantly increases the learning accuracy. Similar to previous tasks, semi-supervised learning can also be performed with crowdsourcing.

Tang and Lease (2011) aim to achieve more accuracy when inferring consensus labels, with correspondingly less labeled training data for estimating worker accuracy by a Naive Bayes approach. We can apply this method in the situation when we have large amount of unlabeled items and a very small set of expert-labeled items.

As human have better learning ability than the algorithm, requesters can provide essential knowledge of how the given task can be performed correctly. This is a well-practiced labeling technique for sophisticated labeling tasks in the data mining field (Sorokin & Forsyth, 2008).

4.4. Sampling

Sampling is in correlation with the selection of a subset with sufficient information so that people can easily verify hypotheses devised from the sample information in the whole datasets. And it is one of the complex tasks in data mining. A challenging problem to be solved here is how the requester should select the appropriate distribution so that benefit of the information gathered from the sample is maximized. Crowds have been proved to be trustworthy data samplers, and they keep working on enhancing the precision of the results in maximally informative samples (Von Ahn, 2009).

4.5. Association rule mining

Data mining techniques have been developed for discovering and identifying underlying association rules among data items. The typical application is shopping baskets analysis. That is, a market analyst can explore relation about which items are purchased together by analyzing purchasing records. However, when referring to human behavior, it is impossible to get access to all the transactions. This is because, typically, the everyday actions of people are not recorded in detail, except in their own memories, which are limited in terms of exact recollection. Indeed, social studies show that instead of full details, people often tend to recall sufficient information in the form of summaries when asked the appropriate questions.

Amsterdamer et al. (2013a, 2013b) lay the foundations of crowd mining for the first time. They define the basic concepts of mining the association rules. Then, they present an integrated system consisted of general-purpose components, incorporating interactive selection of questions to ask, effective mining component, error estimation and so on. Another article from Amsterdamer et al. (2013a, 2013b) present a demo named CrowdMiner. The essence of CrowdMiner is an algorithm enabling the mining of appealing data patterns from the crowd. It allows flexible choice about appropriate questions to ask the crowd as well, with the purpose of gathering more information with fewer questions.

4.6. Validation

Similarly, we can perform task to human to validate the correctness of mining algorithm and predict the mining result of the automated method on large dataset (Barbier et al., 2012).

Agarwal, Liu, Tang, and Yu (2008) want to identify influential bloggers at a blog site. As we know, there is no training and testing data for them to evaluate the efficiency of the proposed model. They use crowdsourced result generated on Digg as a reasonable reference to compare with their automatic techniques. The crowdsourcing results validate their hypothesis.

5. Quality control

Crowdsourcing is a powerful platform for many mining task. Workers may misunderstand the tasks, make mistakes, or deliberately cheat the system, which can cause errors or bad results. Bad answers consume a lot of time and money to filter them out. The reason is twofold.

On one hand, many workers fulfilled tasks to kill time or gain sense of achievement in the beginning, with the payment being only a minor attraction. Nowadays, the overwhelming majority of workers are attracted by the financial reward (Eickhoff & de Vries, 2011). Payment is the easiest method to motivate people. However, monetary incentives can effectively increase participation, but cannot improve quality. As a consequence, a high number of malicious users arise. They try to finish HITs as quickly as possible in order to maximize their profit. This leads to a mass of generic answers.

One the other hand, the worker may lack expertise or skills to handle some kind of complex job (Liu et al., 2012). Incorrect answers may be provided because of this. To tackle this problem, we can provide some basic knowledge to workers before the work, or require some qualifications to prove themselves qualified to finish the task.

In a nutshell, the quality of crowdsourced data has great influence on the mining result, so researchers have to pay careful attention to it. To improve the trustworthiness of mining result, various techniques could be employed. They are discussed as follows, respectively.

5.1. Vote

Voting system hires additional spammers to judge the crowdsourcing outcome, and follows majority rule, which is simple to implement in real world application.

Voting system is quite successful and easy-implement in determining the credibility of messages on the web. In social media sites, people use thumb up or thump down to express their attitudes for or against. For instance, on YouTube, users can provide feedback (positive or negative) for user comments. The website

will hide the comments with too many negative feedbacks automatically (Barbier et al., 2012). Another example is eBay, the buyer vote to seller to give other buyers a reference to the product. And the seller vote according to the buyer's trustworthiness.

Although voting approach has its advantages, we have to admit this approach does have drawbacks. Minority voters have less access to express their views and so researchers would be less likely to benefit from these special ideas.

5.2. Redundant work

Apparently, redundant work means the requester hire redundant workers to finish the same crowdsourcing task. Actually, redundancy work is widely used to identify the correct answers by the requesters.

However, what we should pay attention to is that redundancy is not a panacea. Large-scale redundancy is expensive, and redundant workers sometimes may not lead to good result. Therefore, we can apply redundant workers to controversial item to save money, not to all the items.

5.3. Worker's reputation

Worker's reputation is composed primarily of the worker's accuracy on previously submitted HITs. Reputation is a practical judgment on workers' trustworthiness, which urges people to complete work with high quality continuously. It has become a popular method for evaluation of the quality of workers not only on crowdsourcing platforms, but also on a lot of online forums.

In Amazon Mechanical Turk, requesters can require the level of worker's HITs Approval Rate and some other qualifications, such as language skill. When cheating is detected, the reputation reduces and the system forbids low reputation workers, e.g. users who fail two tasks may be put into blacklist (Heimerl, Gawalt, Chen, Parikh, & Hartmann, 2012).

Allahbakhsh et al. (2012) propose a reputation management framework, which adequately takes into account the values of the tasks completed, the trustworthiness of the assessors, the results of the tasks and the time of evaluation in order to achieve more credible quality metrics for workers and assessors.

Ipeirotis, Provost, & Wang (2010) propose an algorithm that improve the existing advanced techniques of the labeling process in crowdsourcing platform and can be applied when the workers should answer a multiple choice question to complete a task. The algorithm enables the separation of intrinsic error rate from the bias worker. Finally, the algorithm produces a scalar score to measure the intrinsic quality of each worker.

Worker's reputation has potential problems. First, the major determinant factor of human's reputation is the acceptance rate of HITs. The requesters always accept all answers and do not dispose the noisy data right now. Afterwards, requesters do not give feedback to the workers, respectively. The malicious users take advantages of the loophole to receive the increase in reputation and start to complete next tasks. Second, reputation system could not avoid cheating. Ipeirotis (2010) create a scam, which is designed to accelerate the reputation improvement, named rank boosting on his weblog. With this strategy the worker creates a requester account, distributes a number of simple HITs and immediately completes them with his worker account. The worker almost spends no money in boosting his rank.

5.4. Gold standard

Gold standard is the benchmark that is the best available in particular situation. It does not have to be necessarily the best answer for the condition in giving terms. In crowdsourcing domain, we may pay expert to use small group of data to set gold standard, and then we can compare the gold standard data with the HIT's work to judge the reliability and filter out the poor-quality workers.

Bernstein, Teevan, Dumais, Liebling, and Horvitz (2012) build a system on the idea of gold standard questions that the requester has labeled as the ground truth. When encountering the ground truth questions, the worker's answer must include at least one choice from an inclusion list and none from an exclusion list. Le, Edmonds, Hester, and Biewald (2010) insert gold standard data into questions and robustly rejected bad answers to ensure quality when workers made mistakes in those gold standard data. Bernstein, Brandt, Miller, and Karger (2011) extend gold standard questions with novel technique. Also, some trap questions can be mixed with real questions and the system can easily notify the bad answers. Moreover, Callison-Burch and Dredze (2010) suggest that the tasks should be well-priced and make clear enough for workers to follow. They also propose several approaches to restrain cheating, such like using images of sentences instead of text in order to prohibit copying and pasting in translation tasks.

For some situations, like kinds of creative jobs, gold standard seems difficult to set.

5.5. Other solutions for quality control

There are some other solutions for controlling the quality of mining result, which may be more complicated than the methods above, or just the mixture of them.

Liu et al. (2012) design and implement a Crowdsourcing Data Analytics System, CDAS, which is a framework allowing task design and deployment in various crowdsourcing scenarios. Allowing for the human workers' historical performances, the estimation component calculates the accuracy of each generated result. The core part of CDAS commands overall arrangements to process and monitor the human tasks to satisfy user required accuracy.

Liu, Luo, and Li (2013) propose a fancy model, called robust personal classifier (RPC), to improve robustness in crowdsourcing process. The model can create an expertise fraction for each worker automatically, which reflects the intrinsic quality of each worker. The final component of RPC model raises weights for good workers and reduces weights for malicious workers or spammers, which is more proper than same weights for all workers.

Some models and patterns can avoid cheating workers, eliminate irrelevant answers and improve quality. For example, Chilana, Ko, and Wobbrock (2012) identify malicious workers by adding ground truth problem and calculating the differences of global labor rates through an economical model. Chen, Wu, Chang, and Lei (2009) utilize a probabilistic choice model to remove controversial inputs by checking individual consistency and overall consistency of workers. Bernstein et al. (2010) create Find-Fix-Verify crowd programming pattern to separate tasks into three phases to improve the quality. First, the system recruits one set of workers to find underlying areas for improvement. Then it collects a set of possible improvements, and finally filters out incorrect candidates.

As currently employed methods have failed in filtering fraud, Almendra & Schwabe (2009) apply crowdsourcing to improve precision and recall of fraud detection techniques for online trading sites. The experiment showed that workers could distinguish fraudsters from honest sellers precisely and rapidly, according to the personal profiles.

Hirth, Hoßfeld, & Tran-Gia (2010) propose two methods to detect cheating workers based on crowdsourcing: a majority decision (MD) approach and a control group (CG) approach to cross

check the main task. MD is used to eliminate incorrect results. They hire redundant workers to finish the same tasks and compare the results. The majority of the results are considered to be correct. For CG, a single worker works on a main task and a control group consisting of several other workers re-checks the result. There is an assumption that the re-check task has different costs with the main task. Usually the main task is considered to be expensive, while the re-check task is cheap. If the majority of the control group reaches an agreement, the task is supposed to be correctly done.

5.6. Discussion about applicable occasion

In the previous parts, we summarize some tools to control the mining quality. Here we will discuss which quality control mechanism is suitable for which type of mining mechanism.

For classification and clustering tasks, voting and redundant workers are both suitable. And they are easy to implement. If the task needs more field-specific knowledge, we can apply gold standard questions in the process of the crowdsourcing. The gold standard answers are coming from experts in that domain.

For semi-supervised learning, the workers that we need must be able to induct characters from labeled data, which means they have strong ability to learn fast. Hence worker's reputation is a proper reference. The reputation system can reflect the history where this person has done some relevant works before.

For mining the association rule, it is usually about mining the life habits, which are difficult to judge whether they are true or just fake. Voting seems to be an effective way. In this situation, using redundant workers is not a good choice because people differ from behaviors.

The previous experience in working with data mining task on crowdsourcing platform highlights the importance of quality control. Quality control phase can present available information better and serve as an important tool allowing companies and organizations to get what they want most.

6. When not to use crowdsourcing in data mining

The performance of crowdsourcing for data mining is praised in a number of situations, including social hotspot tracking, disaster relief, and homeland security. However, we have to admit that crowdsourcing is not a panacea for solving tasks in all data mining situations. It may lead to poor efficiency and low-quality result if we always adopt the idea of crowdsourcing without thinking over the feasibility. We should consider whether crowdsourcing could achieve better results than traditional algorithms for the current situation before deciding to use crowdsourcing.

Barbier et al. (2012) summarized several scenarios when we cannot use crowdsourcing to solve data mining task. They are listed as follows:

1. The necessity of specific background knowledge. If the background knowledge is too specific and not known by most of people, then using crowdsourcing to do data mining task may not be the proper approach.
2. Improper definition of the problem. Before applying crowdsourcing, we must clarify a single target. If the problem has multiple targets, it's better to separate them into a series of sub-problems.
3. The need of long-term dedication. The most prominent example is software development. Requester could not guarantee that the workers do not leave from the beginning to the end. Crowdsourcing is a highly mobile form which is not suitable for long-term dedication.

7. Future work

There are still many challenges remaining to be solved.

1. Adaptive question delivering system

A real time system that wisely delivers tasks to workers is of great importance. A wise system chooses appropriate next question to ask according to the previous answer that worker provides. It maximums the information gathered from crowd. Not all HITs are equal, and we want to gain more information from the special and "professional" people. Thus different HITs should be allocated different questions according to the level of their knowledge. Adaptive question delivering system targets the problem of reducing uncertainty of the data and saving total cost on both time and money.

2. Recommendation framework

We may design several recommendation frameworks to improve availability of the system. When designing a task, organizations should consider lots of factors, which include, but not limit to, precision, privacy, budget, and priority. Sometimes, a trade-off between these factors is the top concerned problem for the whole mining process. The recommendation frameworks guide requesters to design crowdsourcing task based on their research target and the expected cost.

3. Specific mining algorithms

We need more efforts on specific data mining algorithm design for crowdsourcing. Existing applications and services concentrate on building platforms for crowdsouricng. Algorithms used on crowdsourcing should not be the simple transplant of the ones in being. As far as we know, few efforts have been taken on specific data management and algorithm design for the new data mining process. We should consider the time delay, balance between precision and cost, task design and many other attributes when deliver data mining task on crowdsourcing platform. For example, when posting an image clustering task on the MTurk, it is very common to ask questions in several iterations and wait for feedback from workers. The feedback time is influenced by the task difficulty and complexity. Obviously, more questions will lead to better result. But we intend to ask questions as few as possible to save money. Actually, crowdsourcing itself, as a booming social network application, creates important and precious knowledge during the interactions. And we can use the knowledge to design algorithms that are specialized for the crowdsourcing.

4. New quality guaranteed schemes

The data derived from crowdsourcing process is often noisy and incomplete. Quality must be controlled before working, during working and after working. It is convinced that strict filtering based on task design is a potential method, because the experiment shows the type of the work has great impact on the quality. Creative works will attract less malicious workers. Constant optimization of the worker's reputation system, instead of a simple prior acceptance rate, is another line of thought to protect the quality of crowdsourcing. What is more, a better understanding of human's behavior will be beneficial to improve reliability of the system.

5. Scalability of data mining tasks

Few works has been done to consider the scalability of crowdsourcing for large scale data mining. Actually, the power of

crowdsourcing is fully embodied in managing large datasets. To efficiently harnessing the floods of information will become a great challenge as the scale increases, especially when the information needs to be gathered in some logic sequences.

8. Conclusion

This paper reviews recent work on crowdsourcing-based data mining techniques. Crowdsourcing can do data mining and extract addition information from the datasets more efficiently and intelligently than traditional methods. It has to deal with lots of challenges like the low quality of answers from the crowds to apply crowdsourcing to data mining. In this paper, we point out these challenges and introduce the general procedures of an integrated data mining task in crowdsourcing. The task is often partitioned into three phases: question designing, data mining and quality controlling. We take a deep overview of work in each phase and conclude their contributions.

Besides those discussed in Section 7, there are some promising directions for future research, such as designing crowdsourcing platform especially for governments or companies, algorithms in various steps to process large scale mining task, the background system for requesters to perform detailed analysis. Another future research task is to apply crowdsourcing to discovery knowledge for real expert systems with specific applications.

Acknowledgments

This paper was partially supported by NGFR-China 973 Grant 2012CB316200, NSFC-China Grant 60933001, 61003046, 61111130189 and NGFR-China 863 Grant 2012AA011004.

References

Agarwal, N., Liu, H., Tang, L., & Yu, P. S. (2008). Identifying the influential bloggers in a community. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 207–218). ACM.

Allahbakhsh, M., Ignjatovic, A., Benatallah, B., Beheshti, S. M. R., Bertino, E., & Foo, N. (2012). Reputation management in crowdsourcing systems. In *2012 8th international conference on collaborative computing: networking, applications and worksharing (CollaborateCom)* (pp. 664–671). IEEE.

Almendra, V., & Schwabe, D. (2009). Fraud detection by human agents: A pilot study. In *E-commerce and web technologies* (pp. 300–311). Berlin Heidelberg: Springer.

Alonso, O., & Lease, M. (2011). 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). ACM.

Amsterdamer, Y., Grossman, Y., Milo, T., et al. (2013b). CrowdMiner: Mining association rules from the crowd. *Proceedings of the VLDB Endowment, 6*(12).

Amsterdamer, Y., Grossman, Y., Milo, T., & Senellart, P. (2013a). Crowd mining. In *Proceedings of the 2013 international conference on management of data* (pp. 241–252). ACM.

Barbier, G., Zafarani, R., Gao, H., Fung, G., & Liu, H. (2012). Maximizing benefits from crowdsourced data. *Computational and Mathematical Organization Theory, 18*(3), 257–279.

Bernstein, M. S., Brandt, J., Miller, R. C., & Karger, D. R. (2011). Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on user interface software and technology* (pp. 33–42). ACM.

Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., et al. (2010). Soylent: A word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on user interface software and technology* (pp. 313–322). ACM.

Bernstein, M. S., Teevan, J., Dumais, S., Liebling, D., & Horvitz, E. (2012). Direct answers for search queries in the long tail. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 237–246). ACM.

Boim, R., Greenshpan, O., Milo, T., Novgorodov, S., Polyzotis, N., & Tan, W. C. (2012). Asking the right questions in crowd data sourcing. In *2012 IEEE 28th international conference on data engineering (ICDE)* (pp. 1261–1264). IEEE.

Brabham, D., Sanchez, T., & Bartholomew, K. (2009). Crowdsourcing public participation in transit planning: preliminary results from the next stop design case. Transportation Research Board.

Bragg, J., & Weld, D. S. (2013). Crowdsourcing multi-label classification for taxonomy creation. In *First AAAI conference on human computation and crowdsourcing*.

Budde, A., & Michahelles, F. (2010). Towards an open product repository using playful crowdsourcing. In G. I. Jahrestagung (Ed.), (Vol. 1, pp. 600–605).

Callison-Burch, C., & Dredze, M. (2010). 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* (pp. 1–12). Association for Computational Linguistics.

Chen, Q., Wang, G., & Tan, C. L. (2012). Web image organization and object discovery by actively creating visual clusters through crowdsourcing. *2012 IEEE 24th international conference on tools with artificial intelligence (ICTAI)* (Vol. 1, pp. 419–427). IEEE.

Chen, K. T., Wu, C. C., Chang, Y. C., & Lei, C. L. (2009). A crowdsourceable QoE evaluation framework for multimedia content. In *Proceedings of the 17th ACM international conference on multimedia* (pp. 491–500). ACM.

Chilana, P. K., Ko, A. J., & Wobbrock, J. O. (2012). LemonAid: Selection-based crowdsourced contextual help for web applications. In *Proceedings of the 2012 ACM annual conference on human factors in computing systems* (pp. 1549–1558). ACM.

Chilton, L. B., Little, G., Edge, D., Weld, D. S., & Landay, J. A. (2013). Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the 2013 ACM annual conference on human factors in computing systems* (pp. 1999–2008). ACM.

Eickhoff, C., & de Vries, A. (2011). How crowdsourceable is your task. In *Proceedings of the workshop on crowdsourcing for search and data mining (CSDM) at the fourth ACM international conference on web search and data mining (WSDM)* (pp. 11–14).

Gao, H., Barbier, G., & Goolsby, R. (2011). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems, 26*(3), 10–14.

Gomes, R. G., Welinder, P., Krause, A., & Perona, P. (2011). Crowdclustering. In *Advances in neural information processing systems* (pp. 558–566).

Goolsby, R. (2010). Social media as crisis platform: The future of community maps/crisis maps. *ACM Transactions on Intelligent Systems and Technology (TIST), 1*(1), 7.

Heimerl, K., Gawalt, B., Chen, K., Parikh, T., & Hartmann, B. (2012). CommunitySourcing: Engaging local crowds to perform expert work via physical kiosks. In *Proceedings of the 2012 ACM annual conference on human factors in computing systems* (pp. 1539–1548). ACM.

Hirth, M., Hoßfeld, T., & Tran-Gia, P. (2010). Cheat-detection mechanisms for crowdsourcing. University of Würzburg, Tech. Rep, 4.

Ipeirotis, P. (2010). Be a top mechanical Turk worker: You need $5 and 5 minutes. <http://behind-the-enemy-lines.blogspot.com/2010/10/be-top-mechanical-turk-worker-you-need.html>.

Ipeirotis, P. G., Provost, F., & Wang, J. (2010). Quality management on Amazon mechanical Turk. In *Proceedings of the ACM SIGKDD workshop on human computation* (pp. 64–67). ACM.

Karger, D. R., Oh, S., & Shah, D. (2011). Iterative learning for reliable crowdsourcing systems. In *NIPS* (pp. 1953–1961).

Kazai, G., Kamps, J., Koolen, M., & Milic-Frayling, N. (2011). Crowdsourcing for book search evaluation: Impact of hit design on comparative system ranking. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval* (pp. 205–214). ACM.

Le, J., Edmonds, A., Hester, V., & Biewald, L. (2010). Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation* (pp. 21–26).

Libert, B., & Spector, J. (2007). *We are smarter than me: How to unleash the power of crowds in your business*. Wharton School Publishing.

Liu, X., Lu, M., Ooi, B. C., Shen, Y., Wu, S., & Zhang, M. (2012). Cdas: A crowdsourcing data analytics system. *Proceedings of the VLDB Endowment, 5*(10), 1040–1051.

Liu, Z., Luo, L., & Li, W. J. (2013). Robust crowdsourced learning. In *2013 IEEE international conference on big data* (pp. 338–343). IEEE.

Markowsky, G. (2013). Crowdsourcing, big data and homeland security. In *2013 IEEE international conference on technologies for homeland security (HST)* (pp. 772–778). IEEE.

Quinn, A. J., & Bederson, B. B. (2011). Human computation: A survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1403–1412). ACM.

Reddit website dedicated to solving the Boston Marathon Bombing, no longer a public site (0000). <http://www.reddit.com/r/findbostonbombers>.

Sorokin, A., & Forsyth, D. (2008). Utility data annotation with Amazon mechanical Turk. *Urbana, 51*(61), 820.

Spenser Ackerman. (2013). Data for the Boston Marathon Investigation Will Be Crowdsourced, Wired, April 16. <http://www.wired.com/dangerroom/2013/04/boston-crowdsourced>.

Tang, W., & Lease, M. (2011). Semi-supervised consensus labeling for crowdsourcing. In *SIGIR 2011 workshop on crowdsourcing for information retrieval (CIR)*.

Tran-Thanh, L., Venanzi, M., Rogers, A., & Jennings, N. R. (2013). Efficient budget allocation with accuracy guarantees for crowdsourcing classification tasks. In *Proceedings of the 2013 international conference on autonomous agents and multi-agent systems* (pp. 901–908). International Foundation for Autonomous Agents and Multiagent Systems.

Venetis, P., & Garcia-Molina, H. (2012). Quality control for comparison microtasks. In *Proceedings of the first international workshop on crowdsourcing and data mining* (pp. 15–21). ACM.

Von Ahn, L. (2009). Human computation. In *46th ACM/IEEE design automation conference, DAC'09* (pp. 418–419). IEEE.

Von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM, 51*(8), 58–67.

Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). Recaptcha: Human-based character recognition via web security measures. *Science, 321*(5895), 1465–1468.

Weaver, A. C., Boyle, J. P., & Besaleva, L. I. (2012). Applications and trust issues when crowdsourcing a crisis. In *2012 21st international conference on computer communications and networks (ICCCN)* (pp. 1–5). IEEE.

Wright, A., Chen, E. S., & Maloney, F. L. (2010). An automated technique for identifying associations between medications, laboratory results and problems. *Journal of Biomedical Informatics, 43*(6), 891–901.

Zhai, Z., Kijewski-Correa, T., Hachen, D., & Madey, G. (2012). Haiti earthquake photo tagging: Lessons on crowdsourcing in-depth image classifications. In *2012 seventh international conference on digital information management (ICDIM)* (pp. 357–364). IEEE. 4chan website dedicated to solving the Boston Marathon Bombing, http://imgur.com/a/sUrnA.