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.