Exploration of Applying Crowdsourcing in Geosciences: A Case Study of Qinghai-Tibetan Lake Extraction

Jianghua Zhao(✉), Xuezhi Wang, Qinghui Lin, and Jianhui Li

Computer Network Information Center, Chinese Academy of Sciences,
No. 4 Zhongguancun South Street, Haidian District, Beijing, China
{zjh, wxz, lqh, lijh}@cnic.cn

Abstract. With the emerging of vast quantities of geospatial data, large temporal and spatial scale of data are used in geosciences research nowadays. As a lot of data processing tasks such as image interpretation are hard to be processed automatically, and the data process workload is huge, crowdsourcing is studied as a supplement tool of cloud computing technology and advanced algorithms. This paper outlines the procedure and methodology of applying crowdsourcing in geoscientific data process. And based on the GSCloud platform, a case study of Qinghai-Tibetan Lake Extraction task has been carried out to explore the feasibility of the application of crowdsourcing in geosciences. By analyzing the case, the paper summarizes the problems and characteristics, and advantages and challenges are also presented at last.

Keywords: Crowdsourcing · Geoscientific data process · GSCloud · Qinghai-Tibetan lake extraction

1 Introduction

Vast quantities of data are becoming available at an ever-accelerating rate and remote sensing technology is transforming geosciences today. When investigating global-scale environmental phenomena, the temporal and spatial scale of remote sensing data used in research is so large that it is beyond which scientists have previously encountered [1]. In order to process large data set in reasonable amounts of time, major innovations have been made in parallel computing, programming framework, distributed computer systems, and cloud computing. And all these powerful techniques have been widely used in geoscience data process now. However, some complex tasks which account for a huge part of the application of remote sensing data in geosciences, such as accurate image geometric correction and image interpretation are still very challenging for computers, and it is hard to be automatically processed [2, 3]. So it is time to find a supplement way for massive geoscientific data process.

As each and every human brain has the capacity to process data and anyone can be a problem solver [4], the way of using the distributed thinking resources is an imaginative answer to the problems described above. Crowdsourcing is such a distributed model where an organization or a firm outsourcing a problem or task to an undefined

S. Guo et al. (Eds.): CollaborateCom 2015, LNICST 163, pp. 329–334, 2016.
DOI: 10.1007/978-3-319-28910-6_32

external group (a crowd) [5]. Unlike outsourcing, crowdsourcing allocates work to a collection of individuals, so it is able to access globally distributed qualified talent, and the workforce is flexible. Moreover, it has lower total costs of recruitment, training, supervision and turnover. So crowdsourcing has gained much popularity in numerous fields.

In this paper, an exploration is made to apply crowdsourcing in geoscientific data process tasks. The rest of the paper is organized as follows. In Sect. 2, some studies for crowdsourcing and its application in geosciences are briefly presented. Section 3 describes the process and methodology of applying expert-based crowdsourcing in geoscientific data process. A case study is presented in Sect. 4. Finally, conclusions are made and the future work is discussed in Sect. 5.

2 Related Work

Recently, a lot of academic and industrial organizations have started researching and developing technologies and infrastructures to apply crowdsourcing in geosciences. In the academic area, a lot of platforms have been built. Geo-Wiki is a web-based geospatial portal with open access to Google Earth. Experts or the public can use the high resolution satellite imagery from Google Earth to train and cross check the calibration and validation of land cover products such as GLC-2000, MODIS, and GlobCover on the web [6]. Virtual Disaster Viewer, a social networking tool, uses crowdsourced analysis of remote sensing imagery for earthquake impact and damage assessment [7]. A lot of IT factory also put great effort to the study of crowdsourcing and many crowdsourcing platforms are emerging, such as Elance, CloudCrowd, freelancer, crowdcontent, CrowdFlower and so on. Among the numerous crowdsourcing marketplaces, Amanzon's mechanical turk (MTurk) is the most widely used. MTurk provides a platform for performing tasks that are self-contained, simple, repetitive, and short ones. The tasks requires little specialized skills and the public are motivated by money [8]. Though these platforms gained great popularity, there are some practical limitations. On the one hand, most crowdsourcing platforms assume that complex work can be divided into relatively small and independent ones. However, as science becomes more open, it is possible that some work may not be easily decomposed into units small enough. On the other hand, the online community of these platforms usually are not evaluated or pre-qualified and may offer sub-optimal solutions. Tasks that require higher skills or expertise cannot be easily solved.

Geosciences is a subject that is not only data-driven, but also need a crowd with specific knowledge to solve most data process problems. As Malone and Neis et al. have observed a long-tail mode, that is major contribution always comes from the top few contributors, in crowdsourcing activities, expertsourcing, which crowdsourcing tasks to "crowd" that is comprised of experts or research scientists is proposed [9, 10]. Zhai et al. consider the most important elements in expertsourcing is high reliability and trustworthiness. So the major contribution should come from expert citizens [11]. Tran-Thanh et al. proposes several challenges existing in expertsourcing, including data quality, monetary reward, and so on [12]. Woolley et al. studied how the composition of a community, such as whether it includes randomly selected members or experts, affects results [13]. Dionisio et al. explored expert-based crowdsourcing and try to overcome some deficiencies of crowdsourcing [14].

From the research listed above, it is obvious that expert-based sourcing has great potential in geosciences. Having been in service by providing massive remote sensing data for the public for almost 8 years, Geospatial Data Cloud (GSCloud), a cloud-based platform has more than 95 thousand users who all work or study in fields related with Geoscience. They form an expert community in which members have knowledge or expertise relevant to geosciences, so much more complex tasks can be solved. With this advantage, an exploration of applying expert-based crowdsourcing in geoscientific data processing work is carried out.

3 Methodology

The expert-based crowdsourcing activity GSCloud launched includes the following steps: task definition and division, recruitment and talents selection, task execution and time control, quality control and result aggregation. Each step is described in detail below.

3.1 Task Definition and Division

A massive data analysis task should be divided into small, manageable microtasks. When dividing a huge task, several points should be noticed:

Firstly, choose tasks that are huge, and should be done with human computation. Those tasks that can be process automatically should take advantage of advanced computing algorithm and technology. Secondly, the size of small tasks should be carefully considered. It should be small so that an individual user can generate accurate result within prescribed time period, but also should large enough so that users feel they are making meaningful contributions to the project. So task division demands high human intelligence and skill level. Thirdly, estimate the workload, and guarantee that the workload is proportional to the money reward we pay for it.

3.2 Recruitment and Talents Selection

When crowdsourcing a task, the most important thing is to attract a meaningful number of users, and thus research scientists who are interested in the tasks and are capable of doing it can be easily located. After the task has been published, those users who are interested in the tasks are asked to fill out the registration form, in which his or her related experiences are required. And the implementation plan for this task should be written in detail, which facilitate the GSCloud staff to select appropriate person for this task. Some tasks require users to upload partial preliminary data processing results to evaluate the user's capability of implementing this task directly. Then the applicants will be interviewed and as to those suitable person, an agreement about intellectual property, and quality assurance is signed.

In order to encourage tasking people to produce accurate results and to avoid situations that someone accepts the task but does not accomplish it in time, each task is assigned to two or more person to process independently. The reward is paid in accordance with the quality of the results. Only those whose results are good enough can get the entire monetary reward. The rest may only get part of the total reward.

3.3 Task Execution and Time Control

As geoscience has the characteristics of massive data, GSCloud have to share the data required in the tasks with the task performers. Nowadays, ftp is used to share the massive data required in the task.

After the task has been allocated, the next important thing is to control the time. As those users who obtain the task are part-time, the length of time to implement the task should be long enough to ensure the completion of the task. In order to avoid procrastination, which in turn will affect the quality, timely and effective communication between GSCloud staff and the tasking people is necessary.

3.4 Quality Control and Result Aggregation

During the task execution time period, several procedures were undertaken to validate the accuracy of results including a detailed quality self-evaluation report which covers all the error-prone aspects, a thorough internal review by different quality inspection staff to identify errors and problems, and comparison with high resolution data results, for example, image interpretation results can be assessed using Google Earth. After all these quality control work, there is a time period for tasking people to modify and improve their results. Then the quality of the results are assessed again until the data results are qualified. When all the small tasks are completed, all the partial results of various users are combined into a final, reliable and complete result.

4 Case Study

To study the potential of expert-based crowdsourcing in geoscientific data process, a task of extracting four period of Qinghai-Tibetan lakes from Landsat images during 1995 to 2010 is carried out in this paper. As Qinghai-Tibetan Plateau covers a vast area, and the study period is long, so the lake extraction is a heavy workload. Moreover, as the area is mountainous, and has intricate physiognomic types, plus the influence of heavy cloud and hill shade, it is really difficult to extract lakes automatically. So this is definitely a human computation problem-solving task.

Qinghai-Tibetan Plateau covers an area of nearly 2.6 million square kilometers, which needs more than 150 Landsat images to cover. The images from September to November are used. Each period is divided into 3 small tasks according to the geographical division, and there are total 10 micro tasks, among which the 2005 period task is required a team to execute, so it is not divided. All the task information is published on the website of GSCloud (www.gscloud.cn). And more than 150 users have signed up to apply the tasks. After screening, 18 individuals and two teams are selected. During the task execution process, the professional team of GSCloud supervise all the process, and evaluate the results through different ways, including manual sampling inspection, using high-resolution remote sensing imagery from Google Earth to compare the results, and comparing parallel efforts of the two person executing the same task. Complete results are aggregated for each period, and all the qualified results are obtained in a period of two months. Figure 1 is the map of Qinghai-Tibetan Lake in 2000.

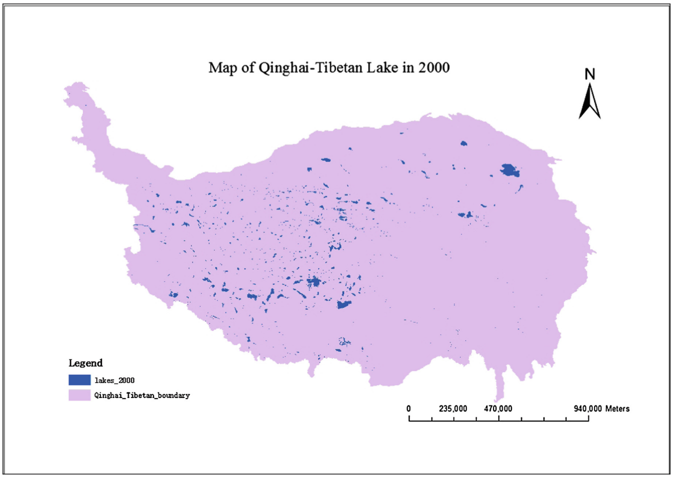

Fig. 1. Map of Qinghai-Tibetan lake in 2000.

In this case, there are some common data quality problems occurred. As the seasonal variations are huge in Qinghai-Tibetan, Landsat images during September to November are normally influenced by cloud and snow, so images of other months will be chosen in the lake extraction task, which will lead to some error. Moreover, as different precipitation on different dates results in different borders of lakes, when the adjacent images are of different time period, it is possible that the lake extraction result in the overlapped area will not be consistent. To ensure the consistency of the lake results in the later data analysis, the larger border of the lakes are kept in the final results. Another problem is the disturbance of snow and ice. So cloud and ice removal work is asked to be done in the data preprocessing. Those areas that are hard to extract lakes, visual interpretation is needed.

5 Conclusion

Applying crowdsourcing, especially expert-based crowdsourcing in geosciences is studied in this paper. The procedure of crowdsourcing in geosciences is described. And by carrying out the Qinghai-Tibetan Lake Extraction task, the advantages and challenges of crowdsourcing can be obtained. By recruiting part-time experts, crowdsourcing does provide data results of good quality in short time and costs little. However, as there is typically little or no prior knowledge about the applicants, how to find appropriate talents or experts for specific tasks remains a challenge. Moreover, the quality requirements of the task should be made clear enough, which demands high professional expertise.

There are a number of directions we are exploring for future work. Most immediately, we are developing a platform, which can be used as collaborative community not only for people to communicate and share knowledge, but also for the public to participate to validate crowdsourced results. Looking further ahead, we are interested in integrating games in the crowdsourcing task to make it more attractive. In addition, more incentive mechanisms will be studied to attract and maintain a pool of experts.

References

1. Bryant, R., Randy, H.K., Lazowska, E.D.: Big-data computing: creating revolutionary breakthroughs in commerce, science and society, pp. 1–15 (2008)
2. Von Ahn, L.: Human computation. In: 46th ACM/IEEE Design Automation Conference, DAC 2009. pp. 418–419. IEEE (2009)
3. Lofi, C., Selke, J., Balke, W.-T.: Information extraction meets crowdsourcing: a promising couple. Datenbank Spektrum **12**(2), 109–120 (2012)
4. Kanefsky, B., Barlow, N.G., Gulick, V.C.: Can distributed volunteers accomplish massive data analysis tasks. In: Lunar and Planetary Science, vol. 1 (2001)
5. Howe, J.: The rise of crowdsourcing. Wired Mag. **14**(6), 1–4 (2006)
6. Fritz, S., et al.: Geo-Wiki: an online platform for improving global land cover. Environ. Modell. Softw. **31**, 110–123 (2012)
7. Barrington, L., et al.: Crowdsourcing earthquake damage assessment using remote sensing imagery. Ann. Geophys. **54**(6), 680–687 (2012)
8. Little, G., et al.: Turkit: tools for iterative tasks on mechanical turk. In: Proceedings of the ACM SIGKDD Workshop on Human Computation. ACM (2009)
9. Malone, T.W., Laubacher, R., Dellarocas, C.: Harnessing crowds: mapping the genome of collective intelligence (2009)
10. Neis, P., Zielstra, D., Zipf, A.: The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007-2011. Future Internet **4**, 1–21 (2012)
11. Zhai, Z., et al.: Expert-citizen engineering: crowdsourcing skilled citizens. In: 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing (DASC). IEEE (2011)
12. Tran-Thanh, L., et al.: Efficient crowdsourcing of unknown experts using multi-armed bandits. In: European Conference on Artificial Intelligence (2012)
13. Woolley, J, Madsen, T.L., Sarangee, K.: Crowdsourcing or Expertsourcing: Building and Engaging Online Communities for Innovation? (2015)
14. Dionisio, M., Fraternali, P., Harloff, E., Martinenghi, D., Micheel, I., Novak, J., Zagorac, S.: Building social graphs from images through expert-based crowdsourcing. In: Proceedings of the International Workshop on Social Media for Crowdsourcing and Human Computation, Paris (2013)