JIMMA UNIVERSITY INSTITUTE OF TECHNOLOGY

FACULTY OF COMPUTING

Discussion Point: Crowd Computing in Data Mining/ Data Analysis

Name	ID
1. Ebrahim Yesuf	Ru0771/08
2. Nasir Abdella	Ru0471/08

Jimma, Ethiopia

Date May 15, 2019

Role of Crowd Computing In Data Mining

Introduction

Crowd computing is a form of distributed work where tasks that are hard for computers to do, are handled by large numbers of humans distributed across the internet. It is an overarching term encompassing tools that enable idea sharing, non-hierarchical decision making and utilization of "cognitive surplus" - the ability of the world's population to collaborate on large, sometimes global projects. Crowd computing combines elements of crowdsourcing, automation, distributed computing, and Mach SSS in learning.

Many data mining tasks cannot be effectively solved by existing machine-only algorithms, such as image classification sentiment analysis and opinion mining For example, given a set of pictures of famous places of interest in the world; we want to cluster them according to the country they belong to Human can use their knowledge to categorize the pictures into countries like "China" or "America", but it is rather hard for machines. Fortunately, crowd computing has been emerged as an effective way to address such machine-hard tasks by utilizing hundreds of thousands of ordinary workers (i.e., the crowd). Thanks to the public crowdsourcing platforms, e.g., Amazon Mechanical Turk (AMT) and Crowd Flower, the access to the crowd becomes easier.

Short Description

Crowd computing and human computation are useful in a number of real-world applications. Crowds generate large data sets useful for natural language process and computer vision; they work together to formulate intelligent responses far beyond what we can automate; and the power intelligent interactive systems currently impossible with automated approaches alone. Crowd Computing is, fundamentally, a distributed computing framework where a big non-trivial task is divided into numerous independent atomic tasks that are distributed over multiple computing devices for processing. These atomic tasks are sometimes referred to as micro-tasks which are kept ready in a job pool. Available crowd worker are being searched for and a set of suitable crowd workers is selected.

Each micro-task from the job pool is assigned to a different crowd worker from that set (though sometimes the same task can be given to different crowd worker to maintain reliability). These micro-tasks are given as simple programs to the crowd worker without any contextual information. After execution of these independent micro tasks, each crowd worker submits the output to the centralized master where all the micro-results are gathered, checked for errors and assembled to get the final result. The other concept is **Data Mining.** Data Mining is a process that uses various techniques to discover hidden relevant information (knowledge or useful patterns) from heterogeneous &distributed historical data stored in large databases, warehouses &other massive information repositories. Data mining intercut with different fields.

For Example

- Machine learning
- Statistics
- Artificial intelligence
- Databases
- Visualization

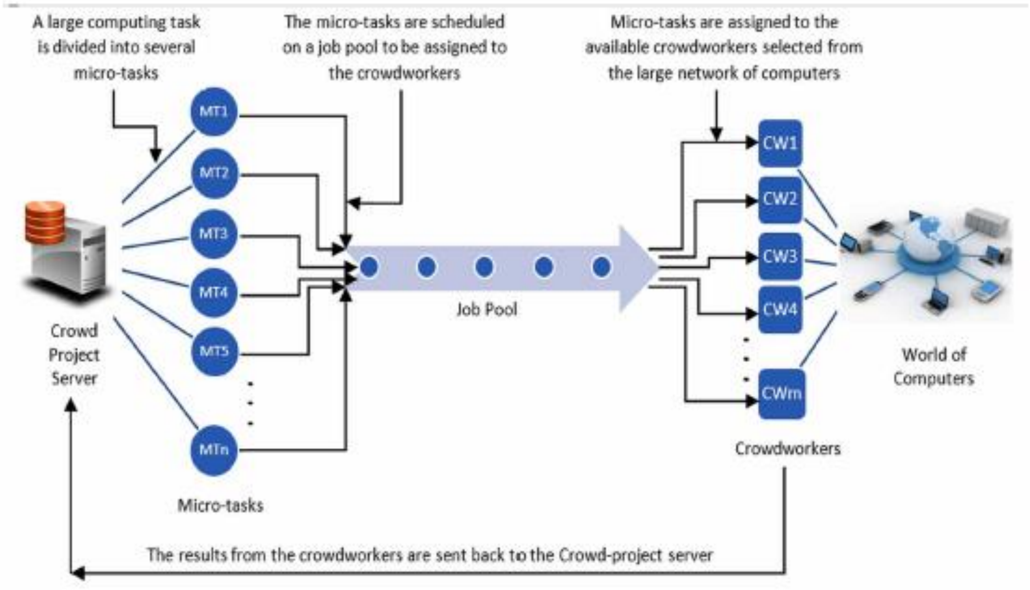
Be for we look rule of crowd computing in data mining, as we know many data mining tasks cannot be completely addressed by automated processes, such as sentiment analysis and image classification. Crowd computing is an effective way to harness the human cognitive ability to process these machine-hard tasks. Thanks to public crowd computing platforms, e.g., Amazon Mechanical Turk and Crowd- Flower, we can easily involve hundreds of thousands of ordinary Workers (i.e., the crowd) to address these machine-hard tasks. Next to this, we look role of crowd computing in data mining process or data mining operations including classification, clustering, pattern mining, machine learning using the crowd (including deep learning, transfer learning and semi-supervised learning) and knowledge discovery. Many data mining tasks cannot be effectively solved by existing machine-only algorithms, such as image classification sentiment analysis and opinion mining. There are many studies that utilize the crowd computing to address the data mining tasks crowd Computing model is expected to take the economics of computing to the next level by drastically enhancing the efficiency of distributed computing resources. **They differ from each other in several ways, as mentioned below**.

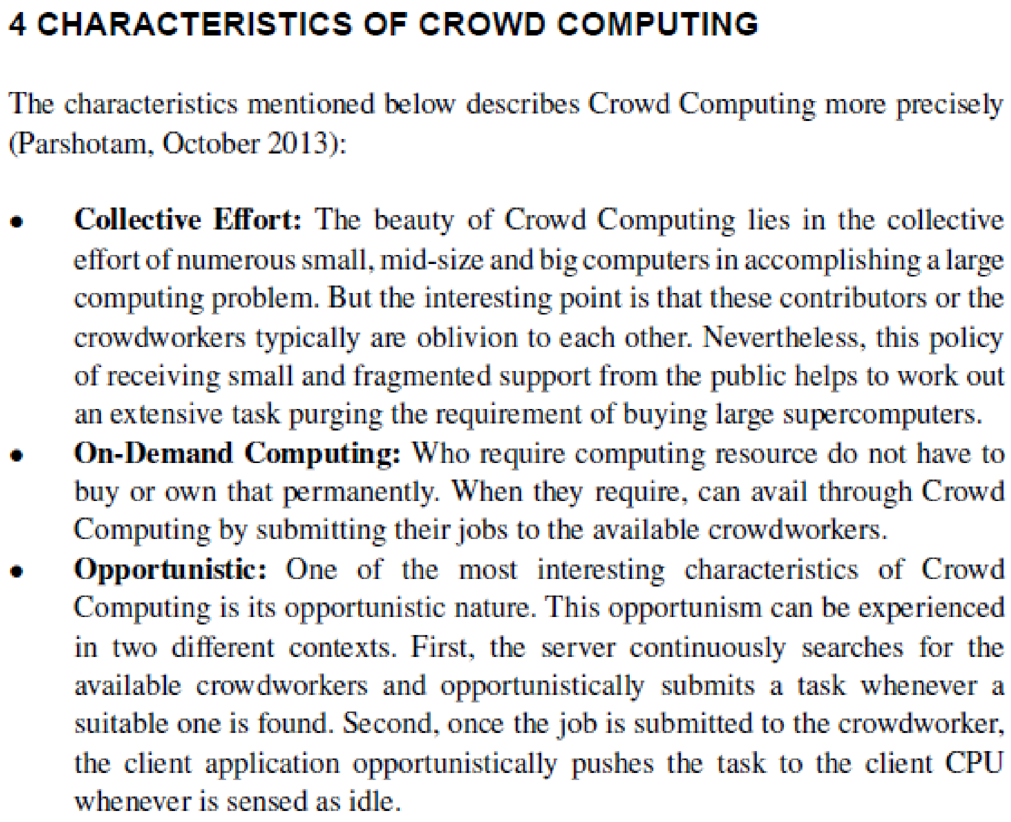
- In cloud computing, the public is the receiver of the service whereas, in Crowd Computing, it is exactly the opposite.
- Cloud resources are centralized but in Crowd Computing, it is highly distributed.
- SLAs (Service Level Agreement) are precisely defined in cloud computing. In volunteer Crowd Computing, SLA may not be of much importance but for non-volunteer Crowd Computing where SLA is a must, there is no standard SLA model.
- Anytime resource availability is guaranteed in cloud computing.

- Crowd Computing facility can be availed without (in case of volunteered service) or with minimal cost. But availing cloud computing always entails money and the cost depends on the type of application and the usage duration.
- Crowd Computing is generally aimed to solve tasks that require huge computing resource. Large projects opt for Crowd Computing. The project initiators maintain the server, the client application, and the middleware as well. While people access cloud computing facility not only for large tasks, but even for small jobs like word processing.

Crowd computing is, fundamentally, a distributed computing framework where a big non-trivial task is divided into numerous independent atomic tasks that are distributed over multiple computing devices for processing.

Figure 1. Basic layout of Crowd Computing

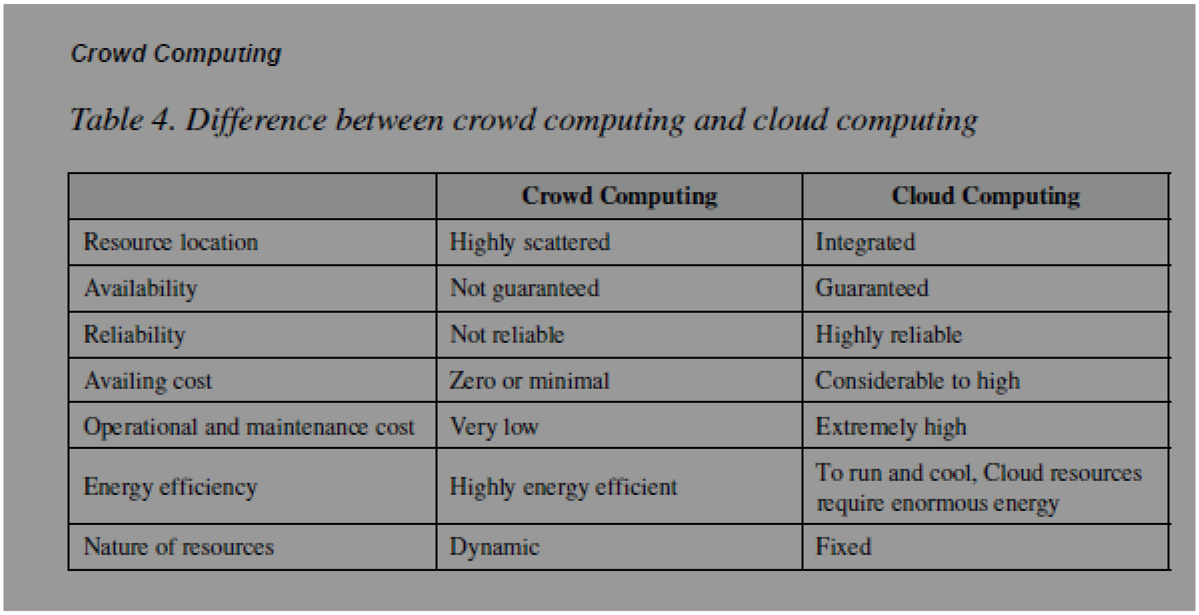

A large computing task is divided into several micro-tasks
The micro-tasks are scheduled on a job pool to be assigned to the crowdworkers
Micro-tasks are assigned to the available crowdworkers selected from the large network of computers
MT1
MT2
MT3
MT4
MT5
MTn
Crowd Project Server
Micro-tasks
Job Pool
CW1
CW2
CW3
CW4
CWm
Crowdworkers
World of Computers
The results from the crowdworkers are sent back to the Crowd-project server



4 CHARACTERISTICS OF CROWD COMPUTING

The characteristics mentioned below describes Crowd Computing more precisely (Parshotam, October 2013):

- **Collective Effort:** The beauty of Crowd Computing lies in the collective effort of numerous small, mid-size and big computers in accomplishing a large computing problem. But the interesting point is that these contributors or the crowdworkers typically are oblivion to each other. Nevertheless, this policy of receiving small and fragmented support from the public helps to work out an extensive task purging the requirement of buying large supercomputers.
- **On-Demand Computing:** Who require computing resource do not have to buy or own that permanently. When they require, can avail through Crowd Computing by submitting their jobs to the available crowdworkers.
- **Opportunistic:** One of the most interesting characteristics of Crowd Computing is its opportunistic nature. This opportunism can be experienced in two different contexts. First, the server continuously searches for the available crowdworkers and opportunistically submits a task whenever a suitable one is found. Second, once the job is submitted to the crowdworker, the client application opportunistically pushes the task to the client CPU whenever is sensed as idle.



Crowd Computing

Table 4. Difference between crowd computing and cloud computing

	Crowd Computing	Cloud Computing
Resource location	Highly scattered	Integrated
Availability	Not guaranteed	Guaranteed
Reliability	Not reliable	Highly reliable
Availing cost	Zero or minimal	Considerable to high
Operational and maintenance cost	Very low	Extremely high
Energy efficiency	Highly energy efficient	To run and cool, Cloud resources require enormous energy
Nature of resources	Dynamic	Fixed

In General, many data mining tasks cannot be completely addressed by automated processes, such as sentiment analysis and image classification. In this description, we will survey and synthesize a wide spectrum of existing studies on crowd computing for data mining. next we review crowd computing in data mining operations, including classification, clustering, pattern mining, machine learning using the crowd (including deep learning, transfer learning and semi-supervised learning) and knowledge discovery. Finally, we provide the emerging challenges in crowdsourced data mining.

Crowd computing is applied in the following tasks

A. **Crowd computing in Pattern Mining:** Crowd pattern mining tries to learn and observe significant patterns based on workers' answers. The problem of discovering significant patterns in crowd's behavior is an important but challenging task. For example, a health researcher is interested in analyzing the performance of traditional medicine and she tries to discover the association rules such that "Garlic can be used to treat flu". In this case, she can neither count on a database which only contains symptoms and treatments for a particular disease, nor ask the healers for an exhaustive list of all the cases that have been treated. So the crowd pattern mining aims to collect the personal rules from crowd workers, aggregate them and find the overall important rules (i.e., general trends). Crowd pattern mining typically generates a huge set of frequent patterns without providing enough information to interpret the meaning of the patterns. It would be helpful if we could also generate semantic annotations for the frequent patterns found, which would help us better understand the patterns.

B. **Crowd computing in Classification:** Some classification tasks are rather difficult for machines but easy for the crowd, e.g., image classification, and crowd-powered classification aims to leverage the crowd's intelligence to classify the data. Since the crowd may make mistake, existing works mainly focus on finding the correct classification from noisy crowd answers.

C. **Crowd computing in Machine Learning**: Crowdsourcing can play an important role in machine learning, such as labeling data or debugging the model. There are several challenges of using the crowd in machine learning field. Firstly, when the number of data to be labeled by human is very large, it is expensive for hiring either experts or the crowd. Therefore, we can utilize transfer learning or semi-supervised learning to do the task. Secondly, since the crowd workers are likely to make mistakes, we have to handle the errors. For example, deep learning can automatically tolerant the errors through the network. In this tutorial, we will discuss the usages of crowdsourcing in these advanced machine leaning algorithms in detail.

D. **Crowd computing in Knowledge Discovery:** We have witnessed the booming of largescale and open-accessible knowledge bases (KBs), which contain thousands of millions of real-world entities, categories and relationships. However, despite the impressive size, no KB is complete. For example, KBs miss many entities, especially the long-tail entities. Thus, some existing works utilize crowdsourcing for knowledge base construction and enrichment, and existing studies can be classified into the following categories.

(a) Crowd-powered knowledge acquisition: combine the crowdsourcing with information extraction techniques for knowledge acquisition in order to fill in missing relations among entities in KBs.

(b) Crowd-powered entity collection: some works utilize crowdsourcing to collect entities that are missing in a KB, e.g., collecting all active NBA players.

(c) Crowd-powered knowledge integration: some works focus on integrating multiple KBs or linking entities in KB to external sources (e.g., web tables).

Conclusion

Crowd computing is a form of distributed work that handled by large numbers of humans distributed across the internet. Many data mining tasks cannot be effectively solved by existing machine. Crowd computing and human computation are useful in a number of real-world applications. Crowd Computing is, fundamentally, a distributed computing framework Crowd computing in data mining concept it will support for the listed things. By applying crowd computing in data mining we can improve our system