COMP90050 Advance Database System Final Report
Crowdsourcing Data Management

Group 1 Members

Lihuan Zhang - 945003 - lihuanz@student.unimelb.edu.au
Xu Wang - 979895 - xuwang2@student.unimelb.edu.au
Ziran Gu - 1038782 - zirang1@student.unimelb.edu.au
Junjie Huang - 1016904 - junjhuang@student.unimelb.edu.au

Lecturer: Prof Rao Kotagiri

School of Computing and Information Systems
University of Melbourne
Australia
May 2019

Abstract

Over the past decade, with the development of technology, more and more people choose to use crowdsourcing, integrate the answers of the crowd to solve the problems that ordinary artificial intelligence can't solve, improve the accuracy of the solutions and the efficiency of solving the problem. However, because the data of crowdsourcing is very large, how to integrate the answers of thousands of people into the database and manage the data perfectly is very worthy of attention.

This article will focus on the introduction that what is crowdsourcing, application of crowdsourcing, along with its strengths and weaknesses, then the second part is the crowdsourcing data management: how the huge data is stored, and some unique algorithms which service for its defects.

Contents

1 Introduction

As is well-known, crowdsourcing is greatly popular in the current era, helping to solve problems which are difficult to be solved by a common approach in different fields and industries. Crowdsourcing provides the ability to send information to the crowd, large-scale human input data, and analysis, and then integrate to provide accurate solutions for the first time in need. The crowd provides a steady stream of professional knowledge and ideas to make a huge contribution to the project or cause. Of course, such large data brings new challenges in storage and management, such as data privacy, accuracy assessment, etc. This article will introduce the overview and function of crowdsourcing and the management of crowdsourced data: The applications and functions of crowdsourcing are discussed in section 2. Section 3 introduces the management for crowdsourcing including task allocation, quality control and data privacy. Section 4 analyzes the characteristics of crowdsourcing data, discusses database technologies which could be used to support crowdsourcing applications, and finally came up with a possible appropriate architecture of the storage system for crowdsourcing applications. The conclusion is presented in section 5.

2 Overview of Crowdsourcing

The overview of crowdsourcing, what is crowdsourcing, how to implement the crowdsourcing in the real world and the application of crowdsourcing will be introduced in this section.

2.1 Definition of Crowdsouring

The definition of crowdsourcing can be interpreted as a project, a problem or a task (will be collectively referred to as ‘project’ at the following article) is solved, through the common contribution of a large group of people. People’s contributions are collected in the form of data as crowdsourcing data, and a large number of data providers work together to achieve the same purpose in order to provide the project with the information they have or create related to the project. Most of the data they provide is closely related to the project, so as to continuously improve the project. Simply put, this project is a joint effort of a large group of people who are willing to enrich the data with the crowd. Wikipedia is a classic example for the crowdsourcing. The purpose of the project is to list all existing words or fields and give a detailed explanation. This is a very large and complex project because it needs to contain all the known knowledge of humans and is constantly updated. Therefore, Wikipedia invites all knowledgeable people to update the current project together and people can create or update terms. After the administrator reviews, these updated terms will be displayed in Wikipedia and provided by a large group of data, so that the continuous supply of fresh data can complete it, this is the crowdsourcing used by Wikipedia.

Crowdsourcing is also applied to Google maps. Google maps, as a project to help people find places and plan routes, needs to contain very detailed and up-to-date road information, and it also provides the ability to view the streetscape of many roads, which requires a very large amount of data support. As one of the users using crowdsourcing, Google maps accepts more than one million mappers to continuously improve the map, which is a typical example of a large-scale crowd to complete a project.

In addition, there are too many examples which are using the crowdsourcing. It is not difficult to find that most of the projects that use crowdsourcing are large-scale jobs that require huge amounts of data, so they often need a group of people to complete it together.

2.2 Crowdsourcing Characteristics

2.2.1 Crowdsourcing Database

In the above we explained the specific meaning of crowdsourcing, and in recent years in order to deal with massive data, therefore, CrowdDB (crowdsourcing database) was born. "Recent panels on crowdsourcing in the database community include 'Web 2.0 and Databases' at VLDB-07 and 'Crowds, Clouds, and Algorithms: Exploring the Human Side of Big Data Applications' at SIGMOD-09."(Doan et al. 2011)

Through a large number of research reports, it can be proved that more and more companies use this database to manage the vast amount of information from various data providers.It is not difficult to imagine that it is costly to answer and deal with people's questions only through the computer, and the answer is not accurate. Therefore, the crowdsourcing database allows the use of human resources to improve the project through the efforts of a large group of people, and the problems that need to be solved by human resources are often cannot be solved only by computers, such as marking content, segmenting images, and identifying the content with human emotion and so on. By using crowdsourcing database, the answer is integrated computer and the human's answer, according to the work requirement, the answer is the common solution between the computer and the human, or after the computer first processes and answers it, then the solution by the computer answer will be identified by a large group of humans to ensure the accuracy. Figure 1 is a good illustration of the workflow: firstly, it send questions to the database through "Query", and then the human group and the machine computer solve the problem and give the answer to the database, then the database integrates the answer and returns.

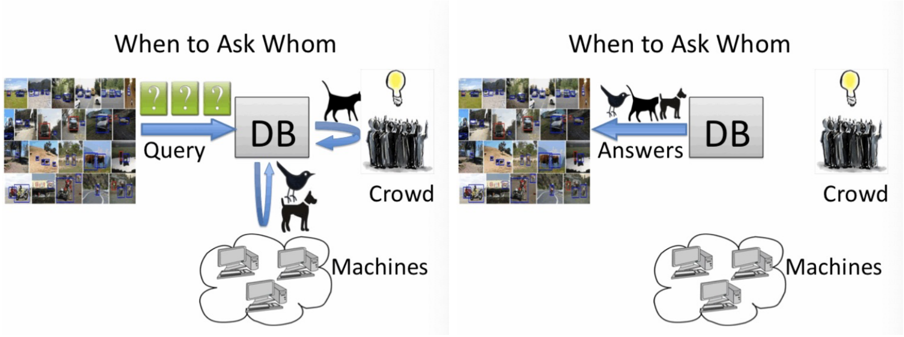

Figure 1: Workflow of crowdsourcing(Mozafari n.d.)

It is essential to use human resources to populate the data in the crowdsourcing database. Although machine learning can handle many problems in the current era, there are still too many answers that it cannot be considered. The crowdsourcing database divides a large project into a very large number of small parts, and each employee process it, then it assembles each employee's answer into a complete solution to return. The benefits of using human resources are numerous, such as the following two points.

- **Finding new data:** Although machine learning is now great smart, it is still difficult to find data which is not currently in the database and to fill and update it. As the example above article, in Wikipedia, it requires human to provide new data constantly to ensure veracity. This is very difficult for machinery and computers,due to it does not identify accurately what is needed for the project, but a humanity can.

- **Comparing data:** After the computer processes the data, it is not difficult to find that there are often some data processing errors. This may be due to the fact that some parts of data is difficult to be completely processed by computer algorithms, and the wrong results cannot be identified by itself, however, humans are good at comparing data. the example include: image annotation, speech recognition, etc., humans can accurately determine the accuracy of data through their own understanding of the data. For example, in image annotation, after the computer processes the image data, the human again analyzes the processed image, and through its own judgment, can determine whether the image is suitable for the answer required by the project, and upload it to the crowdsourcing database to integrate.

CrowdDB is mainly compared with the relational database management system (RDBMS). The CrowdDB has various unique advantages, and one of the most important points is the intervention of human intelligence. As for a project, the user evaluation is essential from the beginning to the end, which means that it can substantially help improve the interaction at first, and the input can act as feedback for the continued improvement (Kittur, Chi, and Suh 2008). For instance, the core of crowdsourcing contains the idea of creating values with users, which means that users or volunteers will participate in and contribute their power for project improvement. As for the working mode, it is different from the ordinary outsourcing mode. The crowdsourcing means that attracting volunteers from outside, instead of using employment way. More details, the strengths are as follows:

- **Low-cost:** As mentioned above, the outsourcing mode emphasizes highly specialized, otherwise, the crowdsourcing model is more focused on the potential of cross-professional innovation, which need not professional, but should be innovative. Thus, people do not need to hire a professional with high salaries when it comes to maintain and improve the crowdsourcing database, which could help decrease the cost of the whole project development. Except for the labor cost reduction, the overhead also can be eliminated efficiently, which means that doing work in-house is suitable to crowdsourcing and it can help decrease the cost that used to create a comfortable working environment or pay for the employees' benefits. Furthermore, using crowdsourcing mode can optimize the management structure, which could limit the number of administrators and workers in one group. Compared with these aspects, it is easy to find that the payment for crowdsourcing is lower than the traditional mode.

- **High efficiency:** The project which is suitable for crowdsourcing to achieve often can be divided into different parts. In this case, the number of volunteers who can simultaneously process the same project online will increase, which can significantly improve the working efficiency. For achieving this goal, the software publisher should smartly assign volunteers to tasks (Roy et al. 2014). For example, when using crowdsourcing database to build a labeling platform with a reasonable distribution mechanism, the goal can be efficiently divided into plenty of parts and uploaded online, everyone can be a volunteer to finish the part in any time, anywhere he/she wants, and the result can be summarized into crowdsourcing database, where the publisher can easily to analyze it. Because of human intelligence, efficiency has been further improved

- **Efficient data collection:** As all we know, training a neural network requires a huge data set, which should contain different factors of the research as much as possible. However, it is impossible for one person to finish the data collection job in a limited time. Therefore, crowdsourcing database plays an important role in this field. For instance, Sigurdsson et al. (2016) discuss using crowdsourcing to collect data for activity understanding. The author sets a standard of human activities, such as the data in-house with a fixed background, and collects data by using crowdsourcing. In other cases, such as building the Wikipedia or Oxford dictionary, crowdsourcing is also useful. Furthermore, building an Open Street Map (OSM), which is an alternative to Google Maps, also encouraged more than one million mappers worked together to supplement detailed street information[1]. Compared with the traditional collecting way, more detailed information of data can be collected by using crowdsourcing mode. Furthermore, in order to ensure the validity of the collected data, the publisher needs to establish a detailed standard in the task description and database establishment.

However, crowdsourcing mode also has some drawbacks which may influence the user experience. On the one hand, although a large group of volunteers can work together at the same time for the same project, sometimes a crowd can return an unpredictable noise that may be of little relevance (Keen 2011) if people do not build a clear data collection standard. To solve the request definition problem, the problem statement formulation, which is defined by Pahl and Beitz (2013) as one of the most important points in the problem-solving processes should be clearly presented, and the publisher team should make an effective way to filter the invalid information. Furthermore, crowdsourcing mode is often used into some non-classified projects, which means that the data collected by volunteers are all insensitive. Use Wikipedia as an example, some data (like entries in Wikipedia) can be easily modified by every user, which means that some of them are incorrect, and even may leak the privacy of other users. Thus, how to protect data privacy is also one of the problems that need to be solved. The solution to building a crowdsourcing database with privacy is going to be discussed in this paper.

On the other hand, some social ethics issues are also widely concerned nowadays. First of all is the standard of remuneration. As mentioned above, companies can use less cost by attracting volunteers to finish the same project than paying for their employees, which is truly a challenge for the professional staff. Originally, employees can get paid for their work by completing tasks, but now they are facing a crisis of unemployment, which will have a huge impact on the traditional labor mode. Furthermore, as Dolmaya (2011) argues, two aspects of translation are should be organized meticulously: one is translation visibility and another one is translation and "minor" languages. Both of them are considered in order to steadily expand the use of crowdsourcing in the world.

[1] https://medium.com/@ico_snovio/crowdsourcing-data-collection-as-a-new-trend-aa37aa4892a

2.2.2 Crowdsourcing platforms

In the past ten years, crowdsourcing has gradually become the mainstream problem solving and data processing method for Internet platforms. Generally speaking, Internet companies will choose to build a crowdsourcing platform to deal with massive problems. The main examples include Mechanical Turk, Mob4hire, uTest, oDesk, Guru, Trada, CloudCrowd, and CloudFlower.

In the following paragraph, Amazon's crowdsourcing platform (Amazon Mechanical Turk (AMT)) will be used as an example to describe the application of crowdsourcing in practical work.

Amazon Mechanical Turk is the Internet crowdsourcing platform owned by Amazon company. Its main job is the same as the crowdsourcing described above, by using human resources to perform work that the computer cannot do, and integrating the answers between all employees and computers to provide the services for Amazon company. Employees who work on the platform can come from anywhere in the world but must have a US credit card. These employees will be tested to verify that they are eligible to complete the task before performing it, and if the results given by the employee are not accurate, the results may be rejected. Since 2006, the platform has grown rapidly. Although Amazon officials have not updated the latest data, more than 200,000 employees have worked for the platform at that time. However, the salary of workers is generally low, if the multiple tasks are quickly executed, the wages of workers per hour are only about 1 US dollar, and only a few cents can be obtained for each small task.

At present, most of the tasks released by the platform are an image or video annotation, writing product descriptions, etc. They divide these large projects into many small tasks called Human Intelligent Task (HIT), which are jobs that workers can complete. For example, identify five photos or videos, write one paragraph of goods description and so on. The requester for each release work must specify how much money is allocated for each HIT, the number and time of employees required to complete each HIT, and the specific requirements of the project. The platform will divide the project requirements submitted by the requester into multiple HITs, determine the number of HITs, and select the appropriate and qualified employees to complete the HIT. Each HIT is often assigned to multiple employees, and when they submit their respective answers, compare and use the most satisfactory answer, then depends on the answer submitted by the employee, pay the employee according to the accuracy.

In this era, there are many companies like Amazon that use crowdsourcing platforms to solve large-scale projects that need to be done manually, Amazon Mechanical Turk is only one typical example. The crowdsourcing platform can well adjust the relationship between labor and computers, make them work together to get a more perfect answer.

2.3 Functions of Crowdsourcing

The main objective of crowdsourcing is how to achieve the human intelligence tasks (HIT) in a low-cost and high-efficiency way, which means that it should be combined with the human knowledge at first, instead of just using computer programming to solve the repetitive problems. As Howe[2] notes the crowdsourcing in his blog, the definition of crowdsourcing is that it often attracts a large group of volunteers to finish one project at the same time, volunteers can choose the published task part as they want and finish it by working in-house. Because of this way of working, people are more selective about how to work with the tasks, which is a much more optional way than the

[2] http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html

original working structure. Finally, the results will be collected into the crowdsourcing database and the publishers can easily analyze the information that what they want. In this paper, the crowdsourcing database is divided into plenty of categories and can be used in different working fields such as solution seeking, data collection, manual tasks, even for customer engagement. In more detail, some function examples are as follow.

- **Crowdsourcing for the labeling system:** There is no doubt that the training effect will be influenced because of the unpredictable noise in the original data set if people ignore labeling in the artificial intelligence and machine learning parts. For instance, if people try to make a neural network to classify the human's faces, then they should label the human's face from pictures as training data, and other background information becomes noise, which should be ignored. Thus, the labeling, such as effective multi-label, etc. Li, M.-L. Zhang, and Geng (2015) plays an essential role, which can cut the useless information and generate basic data set for next training step. However, labeling is truly a repetitive work but still needs human judgment when meeting some complicated data sets like an original photograph or a full video, etc. Thus, using crowdsourcing to build a labeling system is the way to divide the full task into plenty of parts to different volunteers to finish, and Every result is collected and classified online by using the crowdsourcing, which can observably increase the efficiency.

- **Real-time CrowdDB:** CrowdDB can also be built as a real-time way for some situations. For example, a dialog system (in a sense is like a chatbot) is created to make responses the information which comes from the real world by using real-time crowdsourcing (Bessho, Harada, and Kuniyoshi 2012). In real-time CrowdDB, there is not only the data, which is existing in the database but also the volunteer replies (real human beings). It means that people in the task will contribute their power for the response if the system's response time is up to the thread in some extreme cases, which can enhance the user experience. And, the method which is combined with computer programming and human intelligence will significantly enhance work efficiency.

- **Crowdsourcing for the searching system:** In the searching system like Oxford dictionary and Wikipedia, crowdsourcing is just like an optional contribution to the human knowledge for everyone, instead of a compulsive task, which means that people can search the questions which others are eager to know in the searching system, and then they can create or modify the answers for these questions. Use Wikipedia, which is one of the most popular searching systems as the example, it is a website that combines great knowledge that people want to know with high searching efficiency, so that it attracts millions of volunteers to upload their questions or their answers. Because of the crowdsourcing model, each user plays a role of a volunteer, who could help Wikipedia develop its database, and administer only does the daily maintenance operations such as checking if illegal data has appeared. Furthermore, Wikipedia publisher also uses crowdsourcing to develop different kinds of extensions, such as voice reading system, which could help people with reading difficulties and visual impairments search what they want in Wikipedia.

- **Crowdsourcing database for testing:** Testing is an essential part which could verify the system robustness etc. For instance, every online software should be tested before putting into use. However, it is impossible that just let one person find all the bugs in the system because of the mindset. Thus, crowdsourcing is a useful way to build a test platform, which means that gives a place with APIs to volunteers and they could help test the software by using these kinds of APIs and the results could be collected in the background. In that case, it is more like a black box testing that could help programmers find some unpredictable bugs. D. Liu et al. (2012) conduct usability research to test a graduate school's website. From this study, the applicability is clearly evidenced. Comparing the testing result with traditional testing methods, some important usability problems can be found by using crowdsourced testing, which could be a great reference for improving the project.

3 Crowdsourcing Management

3.1 Task Allocation

In space crowdsourcing tasks, task allocation problem refers to the crowdsourcing platform assigns the most suitable crowdsourcing participants for each task according to the geo-location and the task characteristics of crowdsourcing participants. Because different space crowdsourcing applications have different requirements for task assignment, existing studies usually uses 2 different algorithmic models to model different application scenarios. The first one uses a two separate component matching model diagrams to assign a crowdsourcing task to each crowdsourcing workers for a period of time. Its typical application scenario is real-time dedicated vehicle services, such as Uber[3] and DiDi[4] (Jiang et al. 2018). Another one uses task modelling, which assign multiple tasks to crowdsourcing workers at the same time sequentially. For example, Uber Eats[5] and Deliveroo[6]. These two task allocation model are described below.

3.1.1 Matching-based Task Allocation Model

The matching-based model mainly uses maximum or minimum weighted bipartite graph. According to the features in different tasks, the matching-based model can combine the static and dynamic data and calculate the best option for each tasks, choosing out the best crowdsourcing task's participant. In which static data refers to spacial matching and dynamic data refers to the similarity between each tasks. For instance, an Uber Eats' rider can be assigned to another task to Carlton when he/she accepted a task from the Melbourne CBD to Carlton North.

3.1.2 Planning-based Task Allocation Model

This kind of model is suitable for performing multi-crowdsourcing tasks for crowdsourcing participants in a limited time. Therefore, crowdsourcing platforms need to arrange an optimized method for the workers. Similar to the matching-based task allocation model, the planning-based task allocation model can also be divided into static planning model and dynamic planning model according to the time limit of tasks.

[3] https://www.uber.com

[4] https://www.didiglobal.com

[5] https://www.ubereats.com

[6] https://deliveroo.com.au

3.2 Quality Control

Aside from the task allocation issues of crowdsourcing data, the quality of tasks is also an important problem. On one hand, crowdsourcing tasks can be done by many people at the same time, and how to converge them into useful information is worth discussing. On the other hand, crowdsourcing participants may not satisfy people who assign tasks due to the time limitation to each tasks. The common methodology is described below.

3.2.1 Result Convergence

Quality control based on result convergence is a common way to ensure crowdsourcing tasks' quality. Take Uber as an example, result convergence correspond to customer's evaluation after the ride. Specifically, this kind of research usually assigns a group of crowdsourcing participants to each crowdsourcing task, and each participant evaluates the reliability of completing the task correctly according to its historical performance. The core problem of the research is to find N crowdsourcing participants for each crowdsourcing task, and to make the probability that at least one or two crowdsourcing participants can complete the task correctly satisfy the given probability.

3.2.2 Spatiotemporal constraint

The spatiotemporal constraint model is unlike traditional crowdsourcing data management. It usually uses the reliability of crowdsourcing participant set as the core index of quality control, real-time spatiotemporal crowdsourcing data management usually regards the time spent to complete tasks as the key index of quality control. For example, if Uber drivers and passengers are far apart, drivers will have to wait for a long time while drivers have to drive cars with no load.

To sum up, quality control and task allocation affects each other. Quality control provide guidance for task assignments while it also depends on the reasonability of the task allocation strategy.

3.3 Data Privacy

With the improvement of computing and sensing capabilities of mobile internet technology and mobile devices, crowdsourcing based on user's location information emerges as the times require. This kind of location-based crowdsourcing which collects user's geometry information is called spatial crowdsourcing.

Space crowdsourcing, as an exceptional form of crowdsourcing, which employs workers to perform space-related tasks, has been gradually emerging in academia and industry. Typical space crowdsourcing platforms assign space-related missions to adjacent participants, who move to designated locations and complete assigned space tasks. Through space crowdsourcing platforms, people can make better use of group wisdom to complete simple or complex space tasks. Although group wisdom is fully reflected in space crowdsourcing and great benefits can be made, the construction and popularize of the platform are not a easy task. Space crowdsourcing publishes or distributes tasks based on location information submitted by users, while user location information implies sensitive information such as user's identity, home address, and even living habits. In recent years with the continuous development of software and hardware devices, smartphones can act as multi-mode sensors, collecting and sharing assorted types of data, such as images, videos, moving direction and acceleration. Spatial crowdsourcing platforms acquire a large number of user location data through smartphones, which will lead to the leakage of user sensitive information and seriously threaten the privacy and security of users(Zeng et al. 2019). Therefore, three main methods of location privacy protection in space crowdsourcing comes into being.

3.3.1 Differential Privacy Technology

The first one is the protection model based on Differential Privacy Technology (Dwork, Roth, et al. 2014). In 2016, Apple Inc. proposed using differential privacy technology to help discover user behavior from a large amount of data without revealing the privacy of individual users. The definition of differential privacy technology is shown in formula (1). In which ε is the differential privacy of location data(Hassan and Curry 2014). Let random algorithm M to a random query function, Range(M) represents all possible output sets of algorithm M. D_1 and D_2 are two adjacent datasets (that is, D1 and D2 have at most one location record different). S is a subset of Range(M).

$$\ln \frac{Pr[M(D_1) \in S]}{Pr[M(D_2) \in S]} \leq \varepsilon \tag{1}$$

Probabilities $Pr[M(D_1) \in S]$ and $Pr[M(D_2) \in S]$ represents the probabilities of outputs $M(D_1)$ and $M(D_2)$ being S, respectively. The parameter ε is used to measure the intensity of privacy protection. The smaller the parameter ε, the higher the similarity between the two probability density functions, and the higher the intensity of privacy protection. Because of the approximation degree of $Pr[M(D_1) \in S]$ and $Pr[M(D_2) \in S]$ have a direct connection with ε, when the value of ε is appropriate, it is difficult to determine whether the in-situ data set is D_1 or D_2 for a specific output S, and ultimately achieve the purpose of privacy protection.

The differential privacy protection mechanism is a quantifiable, evaluable and provable method based on strict mathematical background. Differential privacy protection technology ensures that sensitive data in original data is protected while maintaining its original statistical properties by adding random noise to the original data to interfere with sensitive personal data, so that analysts can perform benign aggregation analysis. The application of differential privacy technology in spatial crowdsourcing location data publishing can effectively prevent malicious attacks based on background knowledge (L. Zhang, Y. Liu, and Wang 2016).

3.3.2 Spatial Anonymity Technology

In addition to using differential privacy technology to protect the location information of users in spatial crowdsourcing, spatial anonymity technology, which uses a spatial anonymous area to replace the precise location information of users, is also a commonly used privacy protection technology.

K-Anonymity is a privacy technology proposed by Sweeney (2002) for the first time to the disclosure of personal sensitive data. Sweeney (ibid.) proposed that if each record in the data table is at least identical with the other K-1 records in the Quasi-identifier, the data table satisfies K-Anonymity. Privacy protection based on spatial anonymity technology in spatial crowdsourcing usually adopts K-Anonymity model. The definition of K-Anonymity in spatial crowdsourcing location is as follows.

In crowdsourcing, the location attribute of the worker is a Quasi-identifier. In an anonymous space area, the location of any worker in a crowdsourcing space cannot be distinguished from that of at least one other worker. Among them, the Quasi-identifier is the smallest set of attributes, which combines other external information and identifies the location of the target with a greater probability.

For example, the geographic location of a space crowdsourcing worker is A: (-37.7990835,144.9609087), which is the Doug McDonell Building. The geographic location B is not only a point but an area including A, named “The University of Melbourne”. The core idea of spatial anonymity is to expand the location point A into a hidden area B to replace the worker’s accurate location information, in which each worker is hidden among at least K-1 workers. Thus, the attacker can only judge the worker in the concealed area, and can not judge the exact location of the worker in the concealed area.

Traditional research on location data protection mostly uses k-anonymity and l-diversity methods to blur the location information of users, expanding the location of users into a blurred location range, so as to realize the protection of the location information of users. However, ambiguous user location information will lead to lower task allocation efficiency in spatial crowdsourcing. Therefore, the spatial crowdsourcing privacy protection model based on spatial anonymity technology generally achieves fuzzy user location information and guarantees the quality of task allocation by combining other regional partitioning methods.

3.3.3 Data Encryption-based Technology

In addition, few researchers have improved and applied traditional data privacy protection technologies, such as data encryption-based technology, to the space crowdsourcing platform in order to protect the location privacy of space crowdsourcing workers. The principle is to use the Privacy Service Provider (PSP) to store crowdsourcing user’s geolocation information.

4 Data Storage

4.1 Four V’s of Crowdsourcing Application Data (Big Data)

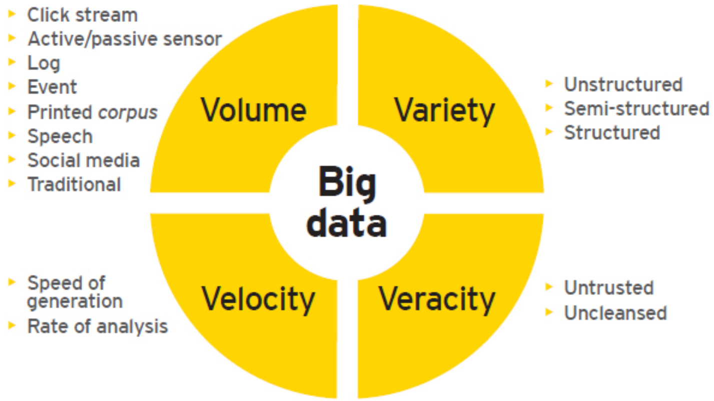

Figure 2: Four V’s of Crowdsourcing Application Data (Big Data)

Basically, crowdsourcing data could be also considered as a form of big data, thus, similar to the features of big data, crowdsourcing applications are also facing four V's challenge: Volume, Variety, Velocity, and Veracity (Foster n.d.), see Figure 2.

- **Volume:** For a crowdsourcing application, like crowdsourcing labeling platforms, the amount of data will rapidly increasing to an unaffordable size (Several EBs) for a single database node, thus it would be necessary to use some distributed database architecture to store all of the data.

- **Variety:** Because the customers of crowdsourcing applications would have different kinds of tasks needed to be solved by the applications, like data cleaning, labeling or even collecting, the databases of these applications usually needed to store various forms of data.

- **Velocity:** Apart from the large volume of data, the velocity of data arrived and be processed are also deserved consideration because the sources of data would keep uploading tasks, thus, for recent successful crowdsourcing platforms, the velocity of concurrent data recorded may reach tens of TB level, thus, it would be a mess without a careful design of the database system.

- **Veracity:** The sources of data are various for crowdsourcing applications and most of the sources are out of control by applications, thus, sometimes the data quality would be low and data inconsistent may be unavoidable. Some data cleaning and quality control procedure should be applied to control the quality.

4.2 Further Characteristics of Crowdsourcing Data

According to the specification of crowdsourcing applications, especially the crowdsourcing labeling platforms, some further characteristics could be abstract from the desired functions.

- **Write $\gg$ Read:** For crowdsourcing applications, the time and disk IO resources consumed by Write request are much more than by Read request. That is because most of the data would be processed, queried and statistics only several times after recorded in the database. Thus the database must support extremely high concurrent writing request. Theoretically, the writing request data throughput may achieve tens of GB level per second.

- **Almost No Update:** Most of the data sent to crowdsourcing applications would never be updated later because usually the data are collected automatically, like images and audios label tasks, the data would only be processed.

- **Unstructured Data:** Similarity to the Variety of data, because the customers may raise various kinds or tasks needed to be solved by the applications, like data labeling, the form of data could be audio, video, text, image or even action. Usually, most of the data are dispersed, there are no obvious relations between them, thus that is almost impossible or need much extra effort to store them in a traditional relational database like MySQL or Oracle.

- **Low Quality:** Similar to the reasons for unstructured data, data sent from different clients, the quality of some data may be low, thus, data cleaning is necessary.

- **Limited Cost:** Due to the large volume of data, if crowdsourcing applications do not control the cost of data storage cost, the cost would increase exponentially, thus, appropriate data storage strategy must be adopted.

- **Data Temperature Strategy:** The data recorded by crowdsourcing application would only be processed in a short time, query and statistics in six months to one year. Thus, the data could be considered as having an additional attribute temperature, which represents the frequency that the data is query or edit, and according to the temperature, the data could be stored with different storage mediums.

4.3 Relative Database Technologies

According to the characteristic, there are some database technologies could be used to support crowdsourcing application data storage.

- **Hadoop Distributed File System (HDFS):** HDFS is a distributed file system, which implemented from the concept of Google File System(GFS), for storing and managing a large volume(PBs) of data. It could be constructed with hundreds or even thousands of nodes and the data sets are stored across the cluster. Firstly, all kinds of files could be stored in HDFS permanently. It could maintain at least three copies of a file automatically, any broken copies would be recovery from other complete copies. Secondly, the files are split into several segments and stored in different nodes and all nodes in cluster related to a file would respond to that query. So, the query could be sped up and the throughput of HDFS could reach an impressive. Finally, the files are only allowed to be read and append but not allowed to be modified(Shvachko et al. 2010).

- **HBase:** HBase is Column-Base Database based on HDFS used to save sparse structured data. Different from the traditional Relational Database, HBase could store inconsistent records and naturally ensure high availability for large data set. HBase is more suitable for the high concurrent write tasks, so it is naturally suitable for crowdsourcing applications. With MapReduce feature, HBase also supports the data cleaning to ensure data quality(George 2011).

- **Multi-Level Storage Strategy:** With the exponentially growing volume of global data, multi-level storage strategy was raised to make a trade-off between the cost of storage and the data access latency. According to the data temperature feature mentioned previously, most of the data would be query and processed frequently in the beginning and would almost never be used again after 1 year. Thus temperature value assigned to data could be generally divided into four levels: hot, warm, cold, and freeze. According to the temperature, data could be stored in different mediums, including Solid State Disk(SSD), Hard Disk(HDD), and Magnetic Tape. The cost of SSD could reach $0.25 per GB, while the cost of HDD remains $0.033 per GB, and the cost of Archive Tape is about $0.02 per GB (Coughlin n.d.). Thus, with a multi-level storage strategy, the cost of storage for crowdsourcing application could be cut 1/3 (Moore et al. 2007).

4.4 Appropriate Architecture

For crowdsourcing application, especially the labeling platforms, the architecture of data storage is significant for the business. According to the characteristics of crowdsourcing data and the database techniques mentioned previously, there is an appropriate architecture for crowdsourcing data storage.

Firstly, the basic service is HDFS, because it naturally supports high concurrent write of unstructured files. The files would be split into segments and recorded by several nodes. Compare with not distributed file systems, the throughput of the whole system may be the same, but the writing speed for a single file is greatly increased to the sum of all related nodes instead of only one machine.

These files only allow being appended instead of modified and the data of crowdsourcing applications also don't need to be updated. Additionally, HDFS maintain three copies of files automatically to ensure data persistence and correctness(Shvachko et al. 2010). On the other hand, HDFS also has some cons, including performing badly when the access latency must within several microseconds, but most of the crowdsourcing applications are not latency sensitive. Thus these features of HDFS totally match the requirements.

Secondly, the files recorded in HDFS should be assigned a "temperature" value to indicate the frequency of being accessed. And then the multi-layer storage strategy could be applied. The temperature of crowdsourcing data should be calculated according to the age of file and access frequency, and the strategy to distinguish temperature depends on companies themselves. For example, Facebook considers the hot data to represent the files recorded within two weeks or accessed every hour, warm data mean the files recorded within 3 months or accessed every day, cold data represent the files recorded within a year or access every month and the freeze data means the files never been used within a year (Gibson n.d.), see Table 1.

Temperature	Age	Access Frequency
Hot	< 2 Weeks	Every Hour
Warm	< 3 Months	Every Day
Cold	≤ 1 Year	Every Month
Freeze	> 1 Year	

Table 1: Data temperature and access frequency

Depend on the temperature of files, different storage medium should be chosen for the files automatically. As HDFS has integrated the ability of multi-layer storage manage, the strategy could be set conveniently with HDFS. What's more, because HDFS maintain at least three copies of a file, copies could be placed in different mediums according to the temperature, instead of storing all three copy in the same medium. For hot data, which are accessed frequently, all three copies could be stored in SSD to speed up the access latency. For warm data, one copy placed in SSD and the rest two placed in HDD to make a trade-off between latency and cost, when querying warm data, the copy store in SSD would respond first and the other two copy would not be accessed until the number of queries increased to a large value. Cool data have two copies in HDD and remain one store in Archive Tape, but the copy in archive tape is only used as a backup when the copies stored in HDD are inconsistent, for example, one of the copies broke, then the HDFS would query the copy in tape to ensure which copy in HDD broke. Freeze data are almost never been used data, thus HDFS could maintain only one copy in HDD, the rest two could be placed in Archive Tapes, then it can maintain the ability to be accessed in a reasonable time and reduces cost as much as possible(Kimmel, Kleiman, and Miller 2015), see Table 2.

Temperature	Copy1	Copy2	Copy3
Hot	SSD	SSD	SSD
Warm	SSD	HDD	HDD
Cold	HDD	HDD	Archive Tape
Freeze	HDD	Archive Tape	Archive Tape

Table 2: Data temperature and copies

The HDFS could migrate data between different mediums automatically, which means the system could assign the temperature value to a data and update it according to the strategy. When data temperature change, the copies would be transmitted to and stored in different mediums.

Finally, HBase could be used to store the information related to the files stored in HDFS, including the address, additional attribution including temperature value, and process results like labels. HBase is base on HDFS thus the HDFS cluster could be used as an HBase cluster directly instead of build another database cluster, save much extra effort. Usually, the results of crowdsourcing data are inconsistent, because data needed to be processed for different tasks for different customers. The performance of HBase is better than other databases for inconsistent data when the data are only allowed to append. What's more, most of the crowdsourcing task needed to be used to do data mining or statistics, thus a column-based database with map-reduce ability is naturally suitable for this kind of analysis.

In conclusion, a possible architecture of the storage system of crowdsourcing applications is constructed with HDFS, HBase, and customized multi-layer storage strategy. With this architecture, crowdsourcing applications could cut down the cost and improve the availability, data safety, and high concurrent throughput of the storage system.

5 Conclusion

There are several key conclusions from this study. This essay gives a general view of the crowdsourcing concepts, techniques, and applications. Firstly, listing some applications of crowdsourcing techniques including CrowdDB, which could return more complete results, and some crowdsourcing platform like labeling platform. Secondly, discuss the functions of crowdsourcing, generally shows how crowdsourcing could be used to benefit current businesses including database, searching engine and wiki-like platform. Thirdly, discuss how to manage the crowdsourcing data including the use K-Anonymous Algorithm to control privacy leaking and came up with a possible storage architecture based on HDFS, HBase and multi-layer storage.

References

Bessho, Fumihiro, Tatsuya Harada, and Yasuo Kuniyoshi (2012). “Dialog system using real-time crowdsourcing and twitter large-scale corpus”. In: *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 227–231.

Coughlin, Tom (n.d.). *The Costs Of Storage*. `https://www.forbes.com/sites/tomcoughlin/2016/07/24/the-costs-of-storage`. Accessed: 1 June 2019.

Doan, A et al. (2011). “Crowdsourcing applications and platforms: A data management perspective”. In: *Proceedings of the VLDB Endowment* 4.12, pp. 1508–1509.

Dolmaya, Julie McDonough (2011). “The ethics of crowdsourcing”. In: *Linguistica Antverpiensia, New Series–Themes in Translation Studies* 10.

Dwork, Cynthia, Aaron Roth, et al. (2014). “The algorithmic foundations of differential privacy”. In: *Foundations and Trends® in Theoretical Computer Science* 9.3–4, pp. 211–407.

Foster, Patrick (n.d.). *What Are the 4 Vs of Big Data & How to Apply Them?* `https://forwardleading.co.uk/blog/What-Are-the-4-Vs-of-Big-Data-How-to-Apply-Them`. Accessed: 1 June 2019.

George, Lars (2011). *HBase: the definitive guide: random access to your planet-size data.* ” O’Reilly Media, Inc.”

Gibson, Dan (n.d.). *Is Your Big Data Hot, Warm, or Cold?* `https://www.ibmbigdatahub.com/blog/your-big-data-hot-warm-or-cold`. Accessed: 1 June 2019.

Hassan, Umair Ul and Edward Curry (2014). “A multi-armed bandit approach to online spatial task assignment”. In: *2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops*. IEEE, pp. 212–219.

Jiang, Yun et al. (2018). “Spatial Crowdsourcing Task Assignment Based on the Quality of Workers”. In: *Proceedings of the 3rd International Conference on Crowd Science and Engineering*. ACM, p. 28.

Keen, Andrew (2011). *The Cult of the Amateur: How blogs, MySpace, YouTube and the rest of today’s user-generated media are killing our culture and economy*. Hachette UK.

Kimmel, Jeffrey S, Steven R Kleiman, and Steven C Miller (2015). *Hybrid media storage system architecture*. US Patent 9,134,917.

Kittur, Aniket, E Chi, and Bongwon Suh (2008). “Crowdsourcing for usability: Using micro-task markets for rapid, remote, and low-cost user measurements”. In: *Proc. CHI 2008*.

Li, Yu-Kun, Min-Ling Zhang, and Xin Geng (2015). “Leveraging implicit relative labeling-importance information for effective multi-label learning”. In: *2015 IEEE International Conference on Data Mining*. IEEE, pp. 251–260.

Liu, Di et al. (2012). “Crowdsourcing for usability testing”. In: *Proceedings of the American Society for Information Science and Technology* 49.1, pp. 1–10.

Moore, Richard L et al. (2007). “Disk and tape storage cost models”. In: *Archiving Conference*. Vol. 2007. 1. Society for Imaging Science and Technology, pp. 29–32.

Mozafari, Barzan (n.d.). *Crowdsourcing Big Data*. `http://istc-bigdata.org/index.php/crowdsourcing-big-data`. Accessed: 30 May 2019.

Pahl, Gerhard and Wolfgang Beitz (2013). *Engineering design: a systematic approach*. Springer Science & Business Media.

Roy, Senjuti Basu et al. (2014). “Optimization in knowledge-intensive crowdsourcing”. In: *arXiv preprint arXiv:1401.1302*.

Shvachko, Konstantin et al. (2010). “The hadoop distributed file system.” In: *MSST*. Vol. 10, pp. 1–10.

Sigurdsson, Gunnar A et al. (2016). “Hollywood in homes: Crowdsourcing data collection for activity understanding”. In: *European Conference on Computer Vision*. Springer, pp. 510–526.

Sweeney, Latanya (2002). “k-anonymity: A model for protecting privacy”. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05, pp. 557–570.

Zeng, Mengjia et al. (2019). “Spatial Crowdsourcing Quality Control Model Based on K-Anonymity Location Privacy Protection and ELM Spammer Detection”. In: *Mobile Information Systems* 2019.

Zhang, Lin, Yan Liu, and RC Wang (2016). “Location publishing technology based on differential privacy-preserving for big data services”. In: *Chin. J. Commun* 37.9, pp. 46–54.