Crowdsensing big data: sensing, data selection, and understanding

Shuying Zhai\textsuperscript{1,a}, Ru Li\textsuperscript{1,b}, Yuange Yang\textsuperscript{1,c}

\textsuperscript{1}Xi’an Mingde Institute of Technology, Fenghe campus of Northwestern Polytechnical University, Xi’an 710124, China
\textsuperscript{a}email: 56211685@qq.com, \textsuperscript{b}email: Liru@mdit.edu.cn, \textsuperscript{c}email: Yangyg@mdit.edu.cn

Abstract. Mobile Crowdsensing (MCS) has become an emerging paradigm for large-scale sensing. It empowers ordinary citizens to contribute data sensed or generated from their mobile devices (e.g., smartphones, wearable devices), aggregates and fuses the data in the cloud for crowd intelligence extraction and human-centric service delivery. The data contributed by the crowd in MCS systems presents the features such as multi-modal, rich-content, spatio-temporal, and human-centric. The key challenges and techniques about crowdsensing big data were discussed. The recent progress of our group in this promising research area was described.

1. Introduction

With the rapid development of embedded devices, wireless sensor networks, Internet of Things, smart mobile terminals, etc., Ubiquitous intelligent systems integrated with sensing, computing and communication capabilities are being widely deployed and gradually integrated into human daily life environment. The ability to calculate and obtain data has also been greatly enhanced. In this context, urban and social sensing has become a frontier research hotspot in the current information field \cite{1,2}.

Generally speaking, urban sensing tasks have the characteristics of wide range, large scale and heavy tasks. The current urban sensing system also mainly relies on pre-installed professional sensing facilities (such as cameras, air detection devices.), which has the problems of limited coverage, high investment and maintenance costs, and the scope of use, objects and application effects are subject to many restrictions. For example, in some places where no cameras are installed, it is difficult to monitor the traffic in that place in real time. In this context, a new sensing mode-mobile crowdsensing (mobile crowdsensing) came into being \cite{3,4}.

Mobile crowdsensing is developed from related concepts such as crowdsourcing and participatory sensing. Crowdsourcing is a professional term invented in 2006 by the American "Wired" magazine to describe a new form of production organization. Specifically, companies/R&D institutions use the Internet to distribute work and use the creativity and capabilities of a large number of users to solve technical problems. Participation sensing was first proposed by researchers at the University of California in 2006 \cite{5}, emphasizing data collection through user participation. In February 2009, Professor Alex Pentland et al. wrote an article in the American "Science" magazine to explain the concept of "Computational Social Science" \cite{6}, and believed that large-scale sensing data can be used to understand individuals, organizations, and society, and it is incompatible with crowdsensing in terms of computing goals. The above several related research directions are based on the participation or data of a large number of users, but emphasize different levels and aspects. In 2012, Professor Liu Yunhao from Tsinghua University merged the above concepts for the first time and proposed the concept of...
"Crowdsensing computing" [3], which uses mobile devices used by a large number of ordinary users as the basic sensing unit, through the Internet of Things/mobile Internet for collaboration. Sensing task distribution and sensing data collection and utilization will finally complete large-scale and complex urban and social sensing tasks. Different from the sensing methods based on the sensor network and the Internet of Things, crowdsensing uses a large number of ordinary users as the source of sensing, emphasizes the use of the mass distribution, flexible mobility and opportunity connectivity for sensing, and provides intelligence for urban and social management Auxiliary support. It can be applied in many important fields, such as intelligent transportation [7], public safety [8], social recommendation [9], environmental monitoring [10], urban public management [11].

In summary, as a brand-new sensing model, Crowdsensing has brought unprecedented opportunities to promote social and urban management innovation. The sensing method that is different from the traditional network also brings many new research questions to it. In particular, the sensing of group intelligence will form multi-modal, rich content, time-space and human-oriented data. Existing models and methods It can't well meet its needs in data processing and understanding, so new calculation models and methods need to be explored. This article takes the group intelligence big data as the target object, expounds and discusses its sensing, optimization and understanding and other key issues, and introduces the author's latest research results in this field.

2. Crowdsensing big data concept and system architecture

Based on the introduction to the background of Crowd big data, this section explains its basic concepts and system architecture. First, introduce the two generation modes and data characteristics of Crowd big data; then, explain the man-machine hybrid intelligence and its cooperation mode; finally, give the typical architecture of Crowd big data system.

2.1. Data generation mode and characteristics

User participation in Crowdsensing is embodied in two modes [12]: offline mobile sensing participation, contributing data through a human-in-the-loop sensing mode; online social networking Media participation, through various mobile social media contribution data, mobile social media can realize the connection of virtual space interaction and physical space elements (such as geographic check-in, activities.). The data contributed by groups through different participation modes is called "group intelligence big data". In the context of the continuous development of information science and technology and its integration into people's daily lives, human behaviour exists in both physical and information spaces. Therefore, Crowd big data increasingly presents the characteristics of extensive group participation, intertwined data time and space, and multi-dimensional target association.

Due to human participation in the data generation process, Crowd big data has many new features compared to traditional perceptual network data. One is that the data of the group intelligence is obtained through a variety of online and offline human participation methods, and is generated in the information space and the physical space at the same time, and due to the role of human beings, different spatial data realize temporal and spatial interweaving and semantic association. Second, the characteristics of human behaviour’s uncertainty and spontaneity make group intelligence data often contain more errors or redundancy, and the quality is uneven, which poses a great challenge to the timely and accurate processing of data. The third is that group intelligence data embodies the integration of people, machines, and things. In the data acquisition process, it also contains a wealth of group intelligence information, such as the interaction characteristics between the group and the perceived object (such as interaction time, location, collection situation, collection mode.). Which provides a foundation for the intelligent integration of humans and machines and efficient data processing.

2.2. Human-machine hybrid intelligence

The research on the combination of human and machine intelligence has a long history. A pioneer in the field of artificial intelligence, Professor Li Clyde of the Massachusetts Institute of Technology published a ground breaking paper in 1960 [13], proposing the idea of "human-computer symbiosis", that is, people
and computers can work together to complete complex tasks. Crowdsensing continues this idea, it tries to solve large-scale sensing and calculation problems through the way of people in the loop. In particular, Crowdsensing makes good use of the complementarity of human and machine intelligence.

- Human intelligence: Humans have learning and reasoning capabilities, such as language, recognition, prediction, and decision-making. At the same time, some individuals or social situations that can be used as input or parameters of Crowdsensing are also called human intelligence, such as social relationships, collaboration, user preferences, and movement patterns. However, human memory and computing speed are limited, the level of data contribution is uneven, and errors or low-quality data are often introduced.

- Machine intelligence: In a narrow sense, machine intelligence is defined as artificial intelligence, that is, the reasoning ability, computing ability, knowledge expression, and problem-solving ability of the machine. Generalized machine intelligence also includes situational information such as the operating status of the system and network connections. However, there are still many problems (such as image annotation, object recognition.) that the machine cannot achieve good performance.

The main feature of Crowdsensing is human participation in the process of solving large-scale problems. However, the capabilities of humans and machines need an effective way to coordinate in order to achieve the enhancement of sensing and computing capabilities. A general Crowdsensing system framework includes three levels of elements, including Crowdsensing at the physical layer, data transmission at the network layer, and data processing at the computing layer. Among them, the Crowdsensing layer uses a large number of mobile or wearable devices to capture the relevant data of the sensing target; the data transmission layer needs to transmit information to the background server through the basic network or opportunistic network; the data processing layer is responsible for ensuring data quality and intelligent analysis and extraction information. At different levels, human and machine intelligence can collaborate in different ways. As shown in Figure 1

![Figure 1. Cooperation between human and machine intelligence](image)

- At the Crowdsensing layer, the computing server is responsible for decomposing complex tasks and assigning them to appropriate groups to complete the tasks. Human participants rely on their knowledge and experience to perform the assigned tasks.
At the data transmission layer, in the Crowdsensing environment, due to the weak network connectivity [3] (such as affected by the uncovered network, user network usage habits.), it is often necessary to reduce the transmission delay of sensing data through opportunistic network connections. The mobility and interactivity of users are the basis for the dynamic connection of opportunistic networks.

At the data processing level, human cognition and expert experience can assist machines to complete data processing tasks more efficiently and accurately, especially when some tasks cannot be solved by existing machine intelligence technology.

2.3. Crowd big data system architecture
A typical Crowd big data system include data source layer, data collection and transmission layer, data processing and calculation layer, and application layer.

- Data source layer: Humans contribute group intelligence data in two ways, namely mobile sensing and mobile social networks. The cloud-end integration method is adopted for data storage and processing, and data storage and calculation tasks can be completed locally or on the server side according to requirements. Access control is an important function of the local end. Users can decide who accesses their data and the scope of access.

- Data collection and transmission layer: A variety of mobile network technologies can be used by group intelligence, including opportunistic networks (such as Bluetooth, Wi-Fi) and infrastructure-based networks (such as 3G/4G). The Crowdsensing network should make data upload transparent to participants and be able to tolerate inevitable network interruptions. In addition, this layer also has functions such as task optimization allocation and incentives for participants.

- Data processing and calculation layer: Machine learning, data mining and reasoning technologies are used to realize the association, integration and understanding of the data contributed by multi-source heterogeneous groups, among which the collaborative computing of human intelligence and machine intelligence is the key content.

- Application layer: Including various applications and services driven by Crowd big data, including social situation awareness, urban computing, environmental pollution monitoring, etc. The author’s previous review article classified and summarized the application of Crowd big data in detail [12].

3. Research challenges and progress
As a new research field, Crowd big data has a series of new research challenges and problems. The progress of these is introduced below.

3.1. Efficient Crowdsensing
Crowdsensing relies on the various sensing and computing capabilities of participating users' mobile terminals for sensing. Compared with traditional sensing networks, participatory sensing nodes have the characteristics of large scale, wide distribution, and complementary capabilities, while tasks have the characteristics of diverse needs, multiple points of concurrency, and dynamic changes. The challenge is how to choose the right participants to efficiently complete the task of city sensing. It is necessary to study the method of selecting participants according to the needs of the perceived task. According to the time and space characteristics of the task, skill requirements and user personal preferences, movement trajectory, movement distance, incentive cost, etc., set optimization goals and constraints and make choices.

3.2. Crowd data optimization
Since different users have a certain overlap in the range of activities, there may be a lot of redundancy in the data collected by Crowdsensing. However, a large number of untrained users as the basic perceptual unit will bring about quality problems such as multi-modal, inaccurate, and inconsistent
perceptual data. The challenge lies in how to achieve high-quality data selection and collection under the circumstances of data redundancy and uneven quality.

3.3. Crowd Data Understanding
How to associate and integrate cross-space multi-source heterogeneous group data and achieve an efficient understanding of sensing goals is another challenge for Crowd big data. Chen S et al. [14] used group trajectory data to construct indoor maps, and filtered abnormal data by mining user access patterns. Cranshaw J et al. [15] extract spatiotemporal features and frequency information from crowdsourced user check-in data to achieve fine-grained friend relationship recognition. Redi M et al. [16] used picture information shared in social networks to portray places of interest. Zheng Y et al. [17] evaluated the blur degree of the group contribution image based on the number of ORB features of the detected image. Researchers from Microsoft Research Asia use mobile phone data for user similarity matching and friend recommendation, and propose an air quality prediction model based on multi-source user contribution data [18].

4. Conclusion
This article expounds the emerging research field of Crowd big data, and describes the characteristics of Crowd data such as cross-space, spatio-temporal interweaving, and human-machine intelligence integration. On this basis, several major research challenges faced by Crowd big data are discussed, including efficient Crowdsensing, group data optimization, and understanding of Crowd data, and the author's research progress in this area is introduced. Crowd big data will play a positive role in promoting the development of human society and information technology, and will promote the harmonious integration of the human-machine-material ternary environment and a large number of "people-centered" applications and service innovations, and more research and technical challenges are also waiting for people to discover and solve.

Acknowledgments
This work was partially supported by the Scientific Research Plan of Ministry of Education of Shaanxi Province (No. 18JK1169) and the Scientific Research Foundation of Mingde Institute of Technology (2019XY01L01).

References
[1] SHETH A. Citizen sensing, social signals, and enriching human experience[J]. IEEE Internet Computing, 2009, 13(4): 87-92.
[2] ZHANG D, GUO B, YU Z. The emergence of social and community intelligence[J]. Computer, 2011, 44(7): 21-28.
[3] LIU Y H. Crowd sensing and computing[J]. Communications of the CCF, 2012, 8(10): 38-41.
[4] GANTI R K, FAN Y, HUI L. Mobile crowdsensing: current state and future challenges[J]. IEEE Communications Magazine, 2011, 49(11): 32-39.
[5] BURKE J A, ESTRIN D, HANSEN M, et al. Participatory sensing[C]// ACM SenSys Workshop on World Sensor Web, October 31-November 3, 2006, Boulder, USA. New York: ACM Press, 2006: 1-5.
[6] LAZER D, PENTLAND A, ADAMIC L, et al. Life in the network: the coming age of computational social science[J]. Science, 2009, 323(5915): 721.
[7] ZHOU P, ZHENG Y, LI M. How long to wait?predicting bus arrival time with mobile phone based participatory sensing[C]// The 10th International Conference on Mobile Systems, Applications, and Services, June 25-29, 2012, Low Wood Bay, UK. New York: ACM Press, 2012: 379-392.
[8] LEE R, WAKAMIYA S, SUMIYA K. Discovery of unusual regional social activities using geotagged microblogs[J]. World Wide Web-internet & Web Information Systems, 2011, 14(4): 321-349.
[9] ZHENG Y, XIE X. Learning travel recommendations from user-generated GPS traces[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(1): 1-29.
[10] MAISONNEUVE N, STEVENS M, OCHAB B. Participatory noise pollution monitoring using mobile phones[J]. Information Polity, 2010, 15(1):51-71.
[11] GUO B, CHEN H, YU Z, et al. FlierMeet: a mobile crowdsensing system for cross-space public information reposting, tagging, and sharing[J]. IEEE Transactions on Mobile Computing, 2015, 14(10): 2020-2033.
[12] GUO B, WANG Z, YU Z W, et al. Mobile crowd sensing and computing: the review of an emerging human-powered sensing paradigm[J]. ACM Computing Surveys, 2015,48(1): 1-31.
[13] LICKLIDER J C R. Man-computer symbiosis[J]. IRE Transactions on Human Factors in Electronics, 1960, HFE-1(1): 4-11.
[14] CHEN S, LI M, REN K, et al. Crowd map: accurate reconstruction of indoor floor plans from crowdsourced sensor-rich videos[C]//2015 IEEE35th International Conference on Distributed Computing Systems, June 29-July 2, 2015, Columbus, USA. New Jersey: IEEE Press, 2015:1-10.
[15] CRANSHAW J, TOCH E, HONG J, et al. Bridging the gap between physical location and online social networks[C]//The 12th ACM International Conference on Ubiquitous Computing, September 26-29, 2010, Copenhagen, Denmark. New York: ACM Press,2010: 119-128.
[16] REDI M, QUERCIA D, GRAHAM L T, et al. Gosling. Like partying? your face says it all. predicting the ambiance of places with profile pictures[C]//The 9th International AAAI Conference on Web and Social Media, May 26-29,2015, Oxford, UK. [S.l.:s.n.], 2015: 295.
[17] ZHENG Y, SHEN G, LI L, et al. Travinavi: self-deployable indoor navigation system[C]//The20th Annual International Conference on Mobile Computing and Networking, September 7-11,2014, Maui, USA. [S.l.:s.n.], 2014: 471-482.
[18] ZHENG Y, LIU F, HSIEH H P U-Air: when ur-ban air quality inference meets big data[C]//The19th ACM SIGKDD International Conference onKnowledge Discovery and Data Mining, August11-14, 2013, Chicago, USA. New York: ACM Press, 2013: 1436-1444.