Digital Technologies, 2015, Vol. 1, No. 1, 39-42
Available online at http://pubs.sciepub.com/dt/1/1/8
© Science and Education Publishing
DOI:10.12691/dt-1-1-8

Pattern-based Data Sharing in Big Data Environments

Muhammad Habib ur Rehman[1,*], Aisha Batool[2]

[1]Faculty of computer science and information technology, University of Malaya, Kuala Lumpur, Malaysia
[2]Department of Computer Science, Iqra University, Islamabad, Pakistan
*Corresponding author: habibcomsats@gmail.com

Received July 16, 2015; Revised July 31, 2015; Accepted August 10, 2015

Abstract The staggering growth in Internet of Things (IoTs) technologies is the key driver for generation of massive raw data streams in big data environments. In addition, the collection of raw data streams in big data systems increases computational complexity and resource consumption in cloud-enabled data mining systems. In this paper, we are introducing the concept of pattern-based data sharing in big data environments. The proposed methodology enables local data processing near the data sources and transforms the raw data streams into actionable knowledge patterns. These knowledge patterns have dual utility of availability of local knowledge patterns for immediate actions as well as for participatory data sharing in big data environments. The proposed concept has the wide potential to be applied in numerous application areas.

Keywords: *big data, edge computing, cloud computing, internet of things*

Cite This Article: Muhammad Habib ur Rehman, and Aisha Batool, "Pattern-based Data Sharing in Big Data Environments." *Digital Technologies*, vol. 1, no. 1 (2015): 39-42. doi: 10.12691/dt-1-1-8.

1. Introduction

The technology landscape is evolving rapidly with the advent of fine-grained sensor rich computing devices (Smartphones, Wearables, IoTs), communication channels (Wi-Fi Direct, IPv6), and cloud computing technologies [1]. The amount of data generated in pervasive, ubiquitous, and big data environments is increasing in size, speed, and variety. Therefore, the challenge of uncovering useful knowledge patterns from these massive stacks of data is increasing day by day. On the other hand, modern digital devices are being empowered by computational and storage facilities [2]. We perceive that the proper utilization of these on-board resources can bring a new life to big data systems.

Mobile data mining systems (MDMS) play a vital role in utilization of on-board resources to uncover knowledge patterns in user locality and provision of local knowledge for immediate local utilization [3]. Existing MDMS are facing the challenges of limited amount of memory, CPU power, battery, and visualization facilities [4]. A thorough literature review exhibits that variety of MDMS were proposed to handle the challenges of context-awareness, energy-efficiency, visualization, local knowledge discovery, and adaptive computations according to on-board available resources [5]. Therefore, all these challenges are needed to be considered while designing the MDMS.

In this paper, we explained big data environments where MDMS could be used as data reducers. Moreover, we explained big data problem in detail and proposed the concept of pattern based data sharing in big data environments. The proposed methodology focuses on reduction of big data in user-facilities instead of directly sharing the data in big data environments. In addition, we discussed the advantages of pattern-based data sharing and its future application areas. The rest of the paper is organized as follows: section-2 discusses related work and section-3 discusses big data problem. In section-4, we discussed the concept of pattern based data sharing and the perceived advantages of proposed methodology are discussed in section-5. In section-6, the perceived advantages are discussed and the article is concluded in section-7.

2. Related Work

Mobile data mining systems are gaining popularity and there are a few well MDMS proposed in recent literature. These MDMS enables knowledge discovery in different local and distributed modes. OMM, StreamAR, and Star utilize local on-board resources in mobile devices to execute knowledge discovery algorithms in light mode and adaptive settings [4,6,7]. The systems collect data streams from on-board and off-board sensors, preprocess data, collect local resource information and adapt their processing behaviors accordingly. OMM, in this case provides a complete toolkit for execution of adaptive and light-weight knowledge discovery algorithms [4]. StreamAR collects accelerometer data and performs online classification in adaptive settings [6]. Star is an extension of StreamAR and it deals with concept drift in online data stream mining mode on accelerometer data [7]. Another variant of local MDMS is mobile WEKA that provides a library of data stream mining algorithms for classification, clustering, and association rule mining in mobile environments [8].

In addition, MDMS also work in distributed settings. These distributed settings include peer-to-peer ad-hoc networks, cloud enabled MDMS or hybrid execution models. PDM is an adaptive light-weight MDMS that works in mobile peer to peer distributed environments [5]. PDM is based on agent-oriented data mining framework where different software agents collaborate to perform collective knowledge discovery in peer-to-peer settings. PDM harnesses granularity-based approaches to execute data mining algorithms in adaptive settings. On the other hand, CARDAP is a cloud-enabled distributed data mining platform where some data mining tasks are performed locally using OMM and other tasks are offloaded to Fog clouds [9] for lateral knowledge discovery [10]. Finally, UniMiner provides a hybrid execution model that performs knowledge discovery in all three modes discussed previously [11]. UniMiner performs local analytics using on-board local resources in context-aware adaptive settings and offloads data mining tasks to other peer devices in locality for collaborative data mining. In case of unavailability of peer devices, UniMiner offloads data mining tasks in cloud environments.A thorough review of literature exhibits that a variety of MDMS could be utilized for pattern based data sharing in big data environments.

3. Big Data Problem

Various big data environments have emerged after the expansion of big data ecosystem in multiple dimensions [12]. These environments vary from in-lab scientific experiments to generation and collection of remote sensory data from heterogeneous data sources. A large number of big data environments have emerged for social media and business analytics systems as well. The massive data production and heterogeneity of data sources has created big data problem (a.k.a. data complexity). This data complexity is due to different v's phenomenon. These v's include volume, velocity, variety, value, veracity, and variability [13].

- Volume: The size of the data produced in big data systems refers to the volume. Generally, a data size which could not be easily processed by conventional systems is known to be volume of big data. Although it terms to be the large data size but it significantly varies in different computational systems. For example, A few GB file is a big data file for a smartphone buy it may not be the case for PC or cloud enabled systems. Similarly, a few TB file is a big data file for a PC but a cloud enabled system may handle it easily. Likewise, few PB size of data set is big data for cloud-based big data systems.
- Velocity: The speed of incoming data streams determines the velocity property of big data system. Velocity is the key challenge in big data systems that increases latency in the big systems. Big data systems handles velocity in two ways: 1) raw data is collected in central data stores for lateral analysis or 2) online data analysis of data streams right after acquisition is performed. In the first approach, big data systems create a delay between data acquisition and knowledge discovery. This strategy is more useful for analysis of historical data. The second approach is more appropriate for real-time data analysis. In this case big data systems need to compromise on the value of overall knowledge patterns due to one-time processing requirements of data streams.
- Variety: The heterogeneity of data sources and data formats in big data systems are represented by the variety property of big data. Data sources for big data systems vary in terms of sensory and non-sensory data sources. In addition, the structured and unstructured data formats also increase variety in big data systems. Big data systems handles the variety challenge effectively to uncover maximum knowledge patterns
- Veracity: The trustworthiness of big data is represented by veracity property of big data. This property is based on the authenticity of data sources and correctness of data. The effective handling of veracity property of big data improves the overall effectiveness of the system.
- Variability: The handling of inconsistencies in big data is attributed with variability. The variable data rates causes computational overhead in peak-load times therefore a proper handling of variability property increases the usefulness of big data systems.
- Value**:** The interestingness of uncovered knowledge patterns is represented by the value property of big data systems. The value is directly affected by other 5V's (velocity, volume, variety, veracity, and variability) therefore a proper balance between all V's brings more value in the big data system. In addition, the effective handling of all other V's is directly proportional to increased value of big data.

3.1. Big Data Complexity

Six V's give the multi-dimensional view of big data hence increase data complexity in big data systems. Data engineers are needed to handle all six V's effectively to increases the overall value of big data systems as well as handle the data complexity. To handle data complexity, the volumes should be reduced in manageable size to effectively and properly apply knowledge discovery algorithms. The velocity should be managed in case of online data analysis so that one-pass data streams algorithms could be applied effectively. The systems must be able to handle the data streams coming from multiple data sources in various formats so that information fusion algorithms could be applied effectively. The maximum and complete data should be acquired from known data sources to handle the veracity and systems must be able to adapt with varying data loads to handle variability of big data. Finally an overall balance between velocity, variety, volume, variability and veracity should be made to harness the maximum value from big data.

4. Pattern-based Data Sharing

Conventionally, most or all of the raw data streams in big data systems are collected at central data stores for lateral analytics. But the increasing growth in MDMS creates an opportunity to handle big data problem and increasing data complexity near the data sources. Here,

mobile device collect, process, and analyze the data streams locally. The resultant knowledge patterns are shared for further correlation analysis in big data systems. The proposed pattern based data sharing workflow is presented in Figure 1.

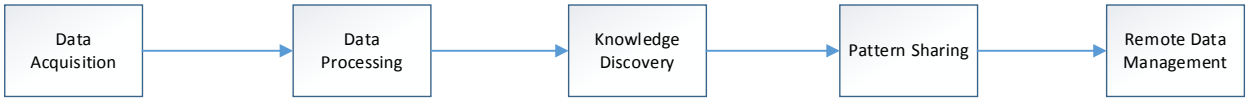

Figure 1. Pattern-based data sharing workflow

The workflow of the proposed strategy is based on five steps for: i) data acquisition, ii) data processing, iii) knowledge discovery, iv) pattern sharing, and v) remote data management. The details of each step are presented below.

i) Data acquisition

The data acquisition is done using on-board sensory and non-sensory data sources. These data sources include accelerometer, gyroscope, GPS sensor, camera, microphone, device-resident data logs, web browser logs, personal contact information, message inbox information and many other data sources. The heterogeneity of data sources and variety of data types bring a wealth of information which, if properly fused and analyzed, it could uncover many interesting knowledge patterns.

ii) Data processing

The data preprocessing and fusion process takes place at this stage. A variety of data preprocessing techniques for feature extraction, feature reduction, noise and dimensionally reduction are applied at this stage. In addition, the data pre-processing techniques, to handle missing values are also handled at this stage. A finely preprocessed data brings more utility in the later stages of pattern sharing workflow.

iii) Knowledge discovery

At this stage the preprocessed data is analyzed using different data mining algorithms for classification, clustering, and association rule mining algorithms. The discovered patterns (classes, clusters, rules) are evaluated using different performance metrics presented in [2]. The interesting patterns are then available for personal usage as well as sharing in big data environments.

iv) Pattern sharing

The interesting patterns are synchronized with remote data stores in cloud environments. The shared patterns are further analyzed in big data environments using different big data mining techniques.

v) Remote data management

The shared patterns are managed in cloud environments using different structured and unstructured data management systems. Here, first the shared patterns are analyzed, integrated and aggregated for further summarization and finally stored in remote data centers.

5. Advantages of Pattern-based Data Sharing in Big Data Environments

The proposed data sharing strategy has some key advantages.

- Local knowledge availability: The execution of local knowledge discovery algorithms enables the provision of local knowledge patterns. Conventionally, user data is collected by big data systems but users do not have the liberty to utilize the data effectively. This strategy brings value to users at personal level as well.
- Complete control over personal data: The proposed data sharing strategy enables users to exercise full control of their personal data patterns. This approach enhances user control and privacy-preservation in big data environments.
- Handles six v's to reduce data complexity: The enablement of knowledge discovery process execution near data sources helps to reduce big data complexity in all six dimensions. For example, the volume of big data is decreased because raw data is processed in local devices and resultant patterns are shared in big data environments. The speed (velocity) of incoming data streams is reduced because of less data transmission from data sources to big data systems. The information about data sources and multiple data types is managed at user devices which effectively handle the variety of big data. The reason is pattern based data sharing which is a uniform format of patterns coming from all data sources. The challenge of variability is met due to less volume, and velocity of incoming data streams. In addition, the issue of veracity is implicitly handled by proposed strategy where missing values and incorrect data are already processed near data sources. Hence the proposed strategy reduces the possibility of unknown data sources and missing/incomplete data. Finally, by overcoming the rest of V's the overall value of big data also increases due to evaluation of crispy patterns instead of raw data streams. The effective hanling of six V's of big data reduce the overall big data complexity.

6. Future Application Areas

The proposed pattern based data sharing strategy could enable many application areas where local knowledge discovery and big data environments are involved. We present three such application areas as follows:

- Mobile Crowd Sensing Applications
  A large number of new applications are being developed, where data from users' personal sensing devices (smartphones and wearable devices) is collected and analyzed for group recommendations. For example, grocery stores collect user's search behavior data from web browser logs and recommends the relevant products accordingly. The proposed strategy in this case will only send grocery items related information to big data environments instead of sending all click stream information. Hence the big data complexity will implicitly be reduced. Similarly, businesses, governments, and other service-providers can collect required relevant information and reduce their computational burden.
- Citizen Participatory Sensing Applications

The evolution in smart-cities and smart-government applications had created potential for participatory data sharing. For examples, smart-city big data systems collecting commuters' trajectories to predict optimal traffic routes or get GPS locations for highly crowded public places for better traffic and law and order management. In this case, the local processing of user trajectories and just sending the travel time from point A to point B not only reduces the communication burden in local mobile devices but also increases user privacy. Conversely, this strategy reduces computational burden of data complexity in big data environments.

- Mobile Social Networks
  Another key application areas involving participatory data sharing is mobile social networks. The proposed strategy helps to reduce communication and computational burden by mining users' physical activities, travel patterns, behaviors, moods, and other personal patterns. The burden of processing in cloud environment automatically reduces because all the patterns are discovered in users' personal sensing devices. In addition, users have complete control over data sharing in mobile social networks. The data patterns shared in MSNs could be further analyzed in big data systems to uncover the hidden patterns on community level.

7. Conclusion

The increasing number of V's in big data introduces big data complexity which increases computational and communicational burdens in big data environments. In this paper we proposed a pattern based data sharing strategy to handle the increasing data complexity. We discussed existing MDMS and the workflow of our proposed strategy. We perceive that our proposed strategy could help to reduce the data complexity. In addition, the availability of local knowledge patterns in user environments enhance the utility of big data systems in communication-efficient way. In future, we will develop a model to effectively harness MDMS for data complexity reduction in big data environments.

References

[1] Gubbi, J., Buyya, R., Marusic, S., and Palaniswami, M.: 'Internet of Things (IoT): A vision, architectural elements, and future directions', Future Generation Computer Systems, 2013, 29, (7), pp. 1645-1660.

[2] Rehman, M.H., Liew, C.S., Wah, T.Y., Shuja, J., and Daghighi, B.: 'Mining Personal Data Using Smartphones and Wearable Devices: A Survey', Sensors, 2015, 15, (2), pp. 4430-4469.

[3] Rehman, M.H., Liew, C.S., and Wah, T.Y.: 'Frequent pattern mining in mobile devices: A feasibility study', in Editor (Ed.)^(Eds.): 'Book Frequent pattern mining in mobile devices: A feasibility study' (IEEE, 2014, edn.), pp. 351-356.

[4] Haghighi, P.D., Krishnaswamy, S., Zaslavsky, A., Gaber, M.M., Sinha, A., and Gillick, B.: 'Open mobile miner: a toolkit for building situation-aware data mining applications', Journal of Organizational Computing and Electronic Commerce, 2013, 23, (3), pp. 224-248.

[5] Gaber, M.M., Gomes, J.B., and Stahl, F.: 'Pocket data mining' (Springer, 2014. 2014).

[6] Abdallah, Z.S., Gaber, M.M., Srinivasan, B., and Krishnaswamy, S.: 'StreamAR: incremental and active learning with evolving sensory data for activity recognition', in Editor (Ed.)^(Eds.): 'Book StreamAR: incremental and active learning with evolving sensory data for activity recognition' (IEEE, 2012, edn.), pp. 1163-1170.

[7] Abdallah, Z.S., Gaber, M.M., Srinivasan, B., and Krishnaswamy, S.: 'Adaptive mobile activity recognition system with evolving data streams', Neurocomputing, 2015, 150, pp. 304-317.

[8] Liu, P., Chen, Y., Tang, W., and Yue, Q.: 'Mobile weka as data mining tool on android': 'Advances in Electrical Engineering and Automation' (Springer, 2012), pp. 75-80.

[9] Bonomi, F., Milito, R., Natarajan, P., and Zhu, J.: 'Fog Computing: A Platform for Internet of Things and Analytics': 'Big Data and Internet of Things: A Roadmap for Smart Environments' (Springer, 2014), pp. 169-186.

[10] Jayaraman, P.P., Gomes, J.B., Nguyen, H.L., Abdallah, Z.S., Krishnaswamy, S., and Zaslavsky, A.: 'CARDAP: A Scalable Energy-Efficient Context Aware Distributed Mobile Data Analytics Platform for the Fog', in Editor (Ed.)^(Eds.): 'Book CARDAP: A Scalable Energy-Efficient Context Aware Distributed Mobile Data Analytics Platform for the Fog' (Springer, 2014, edn.), pp. 192-206.

[11] Rehman, M.H., Liew, C.S., and Wah, T.Y.: 'UniMiner: Towards a unified framework for data mining', in Editor (Ed.)^(Eds.): 'Book UniMiner: Towards a unified framework for data mining' (IEEE, 2014, edn.), pp. 134-139.

[12] Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., and Khan, S.U.: 'The rise of "big data" on cloud computing: Review and open research issues', Information Systems, 2015, 47, pp. 98-115.

[13] Gani, A., Siddiqa, A., Shamshirband, S., and Hanum, F.: 'A survey on indexing techniques for big data: taxonomy and performance evaluation', Knowledge and Information Systems, 2015, pp. 1-44.