A Comprehensive Survey on Semantics for Data Compression

M.V.Gaikwad*, Prof. N.J.Janwe **

[*] Department of computer Engineering, RCERT,Chandrapur
[**]Department of computer Engineering, RCERT,Chandrapur

Abstract- Natural phenomena show that many creatures form large social groups and move in regular patterns. To reduce the data an efficient distributed mining algorithms are used to jointly identify a group of moving objects and discover their movement patterns in wireless sensor networks. In object tracking applications, many natural phenomena show that objects often exhibit some degree of regularity in their movements. To reduce the data volume, various algorithms have been proposed for data compression and data aggregation. In this paper we have surveyed various research papers of movement pattern mining, clustering, and data compression techniques.

Index Terms- Movement pattern Mining,Clustring, Data Compression.

I. INTRODUCTION

RECENT advances in location-acquisition technologies, such as global positioning systems (GPSs) and wireless sensor networks (WSNs), have fostered many novel applications like object tracking, environmental monitoring, and location-dependent service. These applications generate a large amount of location data, and thus, lead to transmission and storage challenges, especially in resource constrained environments like WSNs. To reduce the data volume, various algorithms have been proposed for data compression and data aggregation [1].

Discovering the group movement patterns is more difficult than finding the patterns of a single object or all objects, because we need to jointly identify a group of objects and discover their aggregated group movement patterns. The constrained resource of WSNs should also be considered in approaching the moving object clustering problem. [5]

In order to study the movement behavior of dynamic objects, it is important to take a closer look at movement itself. In other words, it is necessary to know what exactly the variables are that define movement, what constraints and external factors affect movement and most importantly to understand what types of movement patterns can be composed from these primitives of movement.

Generally, movement patterns include any recognizable spatial and temporal regularity or any interesting relationship in a set of movement data, whereas the proper definition (i.e. the instantiation) of “pattern interestingness” depends on the application domain.[2]

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. However, clustering is a difficult problem combinatorially, and differences in assumptions and contexts in different communities has made the transfer of useful generic concepts and methodologies slow to occur[3]

Process of reducing the amount of data needed for storage or transmission of a given piece of information typically by use of encoding techniques. Data compression is characterized as either lossy or lossless depending on whether some data is discarded or not, respectively.

Data compression can reduce the storage and energy consumption for resource-constrained applications.

In [1], Distributed source coding uses joint entropy to encode two nodes’ data individually without sharing any data between them; however, it requires prior knowledge of cross correlations of sources.

The related works are classified into following three types.
Movement Pattern Mining
Clustering
Data compression.

II. MOVEMENT PATTERN MINING

In object tracking applications, many natural phenomenon show that moving objects often exhibit some degree of regularity in their movements. For example, the famous annual wildebeest migration demonstrates that the movement of creatures is temporally and spatially correlated. In addition, biologists have found that many creatures, such as elephants, zebra, whales, and birds, form large social groups when migrating to find food, or for breeding, wintering, or other unknown reasons. These characteristics indicate that the trajectory data of multiple objects may be correlated.Moreover, some research domains, such as the study of animals' social behavior and wildlife migration [16], [17], are more concerned with a group of animals' movement patterns than each individual's. This raises a new challenge of finding moving animals belonging to the same group and identifying their aggregated movement patterns.Many researchers model the temporal-and-spatial correlations of moving objects as sequential patterns in data mining, and various algorithms have been proposed to discover frequent movement patterns [18], [19], [20].

However, such works only consider the movement characteristics of a single object or all objects. Other works, such as [11] take the euclidean distance to measure the similarity of two entire trajectories, and then derive groups of mobile users based on their movement data. Since objects may be close together in some types of terrain, such as gorges, and widely distributed in less rugged areas, such as grassland, their group relationships are distinct in some areas and vague in others. Instead of applying global clustering on entire trajectories, examining partial trajectories of individual areas shows more opportunities of revealing the local group relationships of moving objects.

Following figure shows an example of tracking a group of moving objects

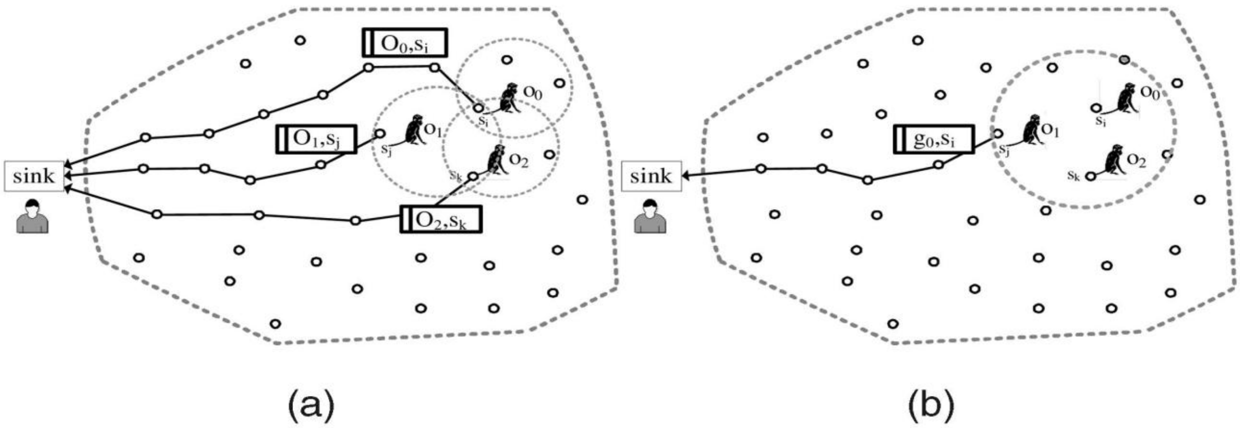

Fig. 12. An example of tracking a group of moving objects.(a) Monitoring multiple objects, respectively. (b) Monitoring multipleobjects with group data aggregation.

Agrawal and Srikant first defined the sequential pattern mining problem and proposed an two algorithms AprioriSome and AprioriAll algorithm to find the frequent sequential patterns. AprioriSome and Apriori- All, have comparable performance, albeit AprioriSome performs a little better when the minimum number of objects that must support a sequential pattern is low.They show that the Accuracy of AprioriSome is better than AprioriAll algorithm[4].

Han et al. consider the pattern projection method in mining sequential patterns and proposed FreeSpan, which is an FP-growth-based algorithm.They reexamin pattern mining problem and proposed novel sequential pattern mining method called Freespan (i,e Frequent pattern- projected sequential pattern mining). The general idea of the method is to integrate the mining of frequent sequences with the frequent pattern and used projected sequence databases to confine the search and the growth of the subsequences fragments.FreeSpan mines the complete set of patterns but greatly reduces the effort of candidate subsequence generation.They show that FreeSpan examines a substantially smaller number of combination of subsequences and runs considerably faster than the Apriori based GSP algorithm.[5]

Yang and Hu developed a new match measure for imprecise trajectory data and proposed TrajPattern to mine sequential patterns. Many variations derived from sequential patterns are used in various applications.They propose the model of the trajectory patterns and a novel measure to represent the expected occurrences of a pattern in a set of imprecise trajectories. The concept of pattern groups is introduced to present the trajectory patterns in a concise manner. Since the Apriori property no longer holds on the trajectory patterns, a new min-maxproperty is identified and a novel TrajPattern algorithm is devised based on the newly discovered property. Last but not least, they also calculate efficiency, and scalability TrajPattern algorithm.[6]

Chen et al. discover path traversal patterns in a Web environment, while Peng and Chen mine user moving patterns incrementally in a mobile computing system. However, sequential patterns and its variations like do not provide sufficient information for location prediction or clustering. First, they carry no time information between consecutive items, sothey cannot provide accurate information for location prediction when time is concerned. Second, they consider the characteristics of all objects, which make the meaningful movement characteristics of individual objects or a group of moving objects inconspicuous and ignored. Third, because a sequential pattern lacks information about its significance regarding to each individual trajectory, they are not fully representative to individual trajectories. [7][8]

Giannotti et al extract T-patterns from spatiotemporal data sets to provide concise descriptions of frequent movements. [9]Tseng and Lin used the TMPMine algorithm for discovering the temporal movement patterns. TMP-Tree for efficiently discovering the temporal movement patterns of objects in sensor networks. This is the first work on mining the movement patterns associated with time intervals in OTSNs. However, novel location prediction strategies that utilize the discovered temporal movement patterns so as to reduce the prediction errors for energy savings. Through empirical evaluation on various simulation conditions and real dataset, TMP-Mine and the prediction strategies are shown to deliver excellent performance in terms of scalability, accuracy and energy efficiency.[10]

III. CLUSTERING

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). clustering denotes the grouping of a set of data items so that similar data items are in the same groups and different data items are placed in distinct groups. Clustering thus constitutes fundamental data analysis functionality that provides a summary of data distribution patterns and correlations in a dataset. Clustering is finding application in diverse areas such as image processing, data compression, pattern recognition, and market research, and many specific clustering techniques have been proposed for static datasets [21].The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. Recently, clustering based on objects' movement behavior has attracted more attention.

A straightforward approach to the clustering of a large set of continuously moving objects is to do so periodically. However, if the period is short, this approach is overly expensive, mainly because the effort expended on previous clustering are not leveraged. If the period is long, long durations of time exist with no clustering information available. Moreover, this brute-force approach effectively treats the objects as static object and does not take into account the information about their movement. For example, this has the implication that it is impossible to detect that some groups of data are moving together. Rather, clustering of continuously moving objects should take into account not just the objects' current positions, but also their anticipated movements.

Figure 2 illustrates the clustering effect where Connected black and the white points denote object positions at the current time and a near-future time.

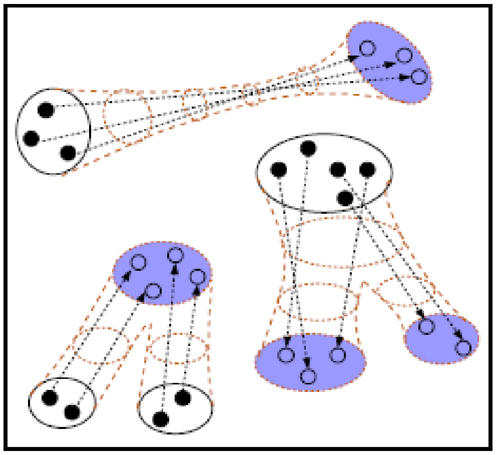

Fig 2: Clustering of moving object

Wang et al. present a new approach to derive groupings of mobile users based on their movement data. They assume that the user movement data are collected by logging location data emitted from mobile devices tracking users. We formally define group pattern as a group of users that are within a distance threshold from one another for at least a minimum duration. To mine group patterns, we first propose two algorithms, namely AGP and VG-growth. In our first set of experiments, it is shown when both the number of users and logging duration are large, AGP and VG-growth are inefficient for the mining group patterns of size two. We therefore propose a framework that summarizes user movement data before group pattern mining. In the second series of experiments, They show that the methods using location summarization reduce the mining overheads for group patterns of size two significantly also they conclude that the cuboid based summarization methods give better performance when the summarized database size is small compared to the original movement database.[11]

Nanni and Pedreschi [12] proposed a density-based clustering algorithm, which makes use of an optimal time interval and the average euclidean distance between each point of two trajectories, to approach the trajectory clustering problem. However, the above works discover global group relationships based on the proportion of the time a group of users stay close together to the whole time duration or the average euclidean distance of the entire trajectories. Thus, they may not be able to reveal the local group relationships, which are required for many applications.

In addition, though computing the average Euclidean distance of two geometric trajectories is simple and useful, the geometric coordinates are expensive and not always available. Approaches, such as EDR, LCSS, and DTW, are widely used to compute the similarity of symbolic trajectory sequences [13], but the above dynamic programming approaches suffer from scalability problem [23]. To provide scalability, approximation or summarization techniques are used to represent original data. Guralnik and Karypis [23] project each sequence into a vector space of sequential patterns and use a vector-based K-means algorithm to cluster objects. However, the importance of a sequential pattern regarding individual sequences can be very different, which is not considered in this work.

IV. DATA COMPRESSION

Data compression can reduce the storage and energy consumption for resource-constrained applications.
Pradhan et al.[1] uses distributed source (Slepian-Wolf) coding technique for data compression which uses joint entropy to encode two nodes' data individually without sharing any data between them; however, it requires prior knowledge of cross correlations of sources.in this work they presented a new domain of collaborative information communication and processing through the framework on distributed source coding. This framework enables highly effective and efficient compression across a sensor network without the need to establish inter-node communication, using well-studied and fast error- correcting coding algorithms.
A. Scaglione and S.D. Servetto [14], used routing algorithm for data compression and also combine data compression with routing by exploiting cross correlations between sensor nodes to reduce the data size.
In [15], a tailed LZW has been proposed to address the memory constraint of a sensor device. Summarization of the original data by regression or linear modeling has been proposed for trajectory data compression.

V. CONCLUSION

With the growth of the data in resource constrined application there is need to compress the amount of data shared between the two nodes.In this survey paper, we briefly explored various techniques of movement pattern mining ,clustering and data compression suggested by authors. Anyway we believe that the most interesting research area deals with the discovering of semantics within movement pattern mining so to improve the results of data compression. Efforts in this direction are likely to be the most fruitful and compresses the much more location data quite effectively.

REFERENCES

[1] S.S. Pradhan, J. Kusuma, and K. Ramchandran, "Distributed Compression in a Dense Microsensor Network," IEEE Signal Processing Magazine, vol. 19, no. 2, pp:51-60, Mar. 2002.

[2] Andrienko, N. & Andrienko, G. Designing Visual Analytics Methods for Massive Collection of Movement Data. Cartographica 2007, 42(2): 117-138.
A.K. Jain,M.N. Murtyand P.J. Flynn Isaac council 1999

[3] R. Agrawal and R. Srikant, "Mining Sequential Patterns," Proc. 11th Int'l Conf. Data Eng., pp. 3-14, 1995..

[4] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. Hsu, "Freespan: Frequent Pattern-Projected Sequential Pattern
Mining," Proc. ACM SIGKDD, pp. 355-359, 2000.

[5] J. Yang and M. Hu, "Trajpattern: Mining Sequential Patterns from Imprecise Trajectories of Mobile Objects," Proc. 10th Int'l Conf. Extending
Technology, pp. 664-681, Mar. 2006.

[6] M.-S. Chen, J.S. Park, and P.S. Yu, "Efficient Data Mining for Path Traversal Patterns," Knowledge and Data Eng., vol. 10, no. 2, pp. 209-221, 1998.

[7] W.-C. Peng and M.-S. Chen, "Developing Data Allocation Schemes by Incremental Mining of User Moving Patterns in a Mobile Computing System," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 1, pp. 70-85, Jan./Feb. 2003

[8] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory Pattern Mining," Proc. ACM SIGKDD, pp. 330-339, 2007.

[9] V.S. Tseng and K.W. Lin, "Energy Efficient Strategies for Object Tracking in Sensor Networks: A Data Mining Approach," J. Systems and Software, vol. 80, no.10, pp. 1678-1698, 2007

[10] Y. Wang, E.-P. Lim, and S.-Y. Hwang, "Efficient Mining of Group Patterns from User Movement Data," Data Knowledge Eng., vol. 57, no. 3, pp. 240-282, 2006.

[11] M. Nanni and D. Pedreschi, "Time-Focused Clustering of Trajectories of Moving Objects," J. Intelligent Information Systems, vol. 27, no. 3, pp. 267-289, 2006.

[12] L. Chen, M. Tamer O ¨ zsu, and V. Oria, "Robust and Fast Similarity Search for Moving Object Trajectories," Proc. ACM SIGMOD, pp. 491-502, 2005.
.

[13] A. Scaglione and S.D. Servetto, “On the Interdependence of Routing and Data Compression in Multi-Hop Sensor Networks,” Proc. Eighth Ann. Int’l Conf. MobileComputing and Networking, pp. 140-147, 2002

[14] C.M. Sadler and M. Martonosi, “Data Compression Algorithms for Energy-Constrained Devices in Delay Tolerant Networks,” Proc. ACM Conf.
Embedded Networked Sensor Systems, Nov. 2006.

[15] C.M. Sadler and M. Martonosi, “Data Compression Algorithms for Energy-Constrained Devices in Delay Tolerant Networks,” Proc. ACM Conf. Embedded Networked Sensor Systems, Nov. 2006.

[16] G. Shannon, B. Page, K. Duffy, and R. Slotow, “African Elephant Home Range and Habitat Selection in Pongola Game Reserve, South Africa,” African Zoology,
vol. 41, no. 1, pp. 37-44, Apr. 2006

[17] M. Morzy, “Prediction of Moving Object Location Based on Frequent Trajectories,” Proc. 21st Int’l Symp. Computer and Information Sciences, pp. 583-592Nov. 2006.

[18] M. Morzy, “Mining Frequent Trajectories of Moving Objects for Location Prediction,” Proc. Fifth Int’l Conf. Machine Learning and Data Mining in PatternRecognition, pp. 667-680, July 2007

[19] V.S. Tseng and K.W. Lin, “Energy Efficient Strategies for Object Tracking in Sensor Networks: A Data Mining Approach,”J. Systems and Software, vol. 80, no 10, pp. 1678-1698, 2007.

[20] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: An efficient data clustering method for very large databases. In *Proc. ACM SIGMOD*, pp. 103–114, 1996.

[21] C.S. Jensen, Dan Lin, and Beng Chin Ooi, “Continuous Clustering of Moving Objects,” IEEE Transl. on Knowledge and Data Engineering,vol. 19, pp. 1161-1174,2007

AUTHORS

First Author – Mayur V.Gaikwad, B.E (I.T), RCERT Chandrapur,mayurgkwd@gmail.com
Second Author – Nitin Janwe, MTECH (CSE), RCERT Chandrapur,nitinj_janwe@yahoomail.com
Third Author – Author name, qualifications, associated institute (if any) and email address.

Correspondence Author – Mayur V.Gaikwad,mayurgkwd@gmail.com,7387650299.