ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Formulate a Hit Item Replacement and Propose Cluster Ensembling (CE) Algorithm for Data Compression

E. Vasantha
Shadan Engineering College, India
vasanthaget@gmail.com

Abstract

In this paper, we first propose an efficient distributed mining algorithm to jointly identify a group of moving objects and discover their movement patterns in wireless sensor networks. Afterward, we propose a compression algorithm, called 2P2D, which exploits the obtained group movement patterns to reduce the amount of delivered data. The compression algorithm includes a sequence merge and an entropy reduction phases. In the sequence merge phase, we propose a Merge algorithm to merge and compress the location data of a group of moving objects. In the entropy reduction phase, we formulate a Hit Item Replacement (HIR) problem and propose a Replace algorithm that obtains the optimal solution. Moreover, we devise three replacement rules and derive the maximum compression ratio. The experimental results show that the proposed compression algorithm leverages the group movement patterns to reduce the amount of delivered data effectively and efficiently.

Keywords: data compression, hit item replacement.

Introduction

In object tracking applications, many natural phenomena show that objects often exhibit some degree of regularity in their movements. For example, the famous annual wildebeest migration demonstrates that the movements of creatures are temporally and spatially correlated. Biologists also have found that many creatures, such as elephants, zebra, whales, and birds, form large social groups when migrating to find food, or for breeding or wintering. These characteristics indicate that the trajectory data of multiple objects may be correlated for biological applications. Moreover, some research domains, such as the study of animals' social behavior and wildlife migration are more concerned with the movement patterns of groups of animals, not individuals; hence, tracking each object is unnecessary in this case. This raises a new challenge of finding moving animals belonging to the same group and identifying their aggregated group movement patterns. Therefore, under the assumption that objects with similar movement patterns are regarded as a group, we define the moving object clustering problem as given the movement trajectories of objects, partitioning the objects into non overlapped groups such that the number of groups minimized. Then, group movement pattern discovery is to find the most representative movement patterns regarding each group of objects, which are further utilized to compress location data. 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. However, few of existing approaches consider these issues simultaneously. On the one hand, the temporal-and-spatial correlations in the movements of moving objects are modeled as sequential patterns in data mining to discover the frequent movement patterns. However, sequential patterns 1) consider the characteristics of all objects, 2) lack information about a frequent pattern's significance regarding individual trajectories, and 3) carry no time information between consecutive items, which make them unsuitable for location prediction and similarity comparison. On the other hand, previous works, such as measure the similarity among these entire trajectory sequences to group moving objects. Since objects may be close together in some types of terrain, such as gorges, and widely distributed in less rugged areas, their group relationships are distinct in some areas and vague in others. Thus, approaches that perform clustering among entire trajectories may not be able to identify the local group relationships. In addition, most of the above works are centralized algorithms which need to collect all data to a server before processing. Thus, unnecessary and redundant data may be delivered, leading to much more power consumption because data transmission needs more power than data processing in WSNs. In we have proposed a clustering algorithm to

ISSN: 2277-9655
Impact Factor: 1.852

find the group relationships for query and data aggregation efficiency. The differences of and this work are as follows: First, since the clustering algorithm itself is a centralized algorithm, in this work, we further consider systematically combining multiple local clustering results into a consensus to improve the clustering quality and for use in the update-based tracking network. Second, when a delay is tolerant in the tracking application, a new data management approach is required to offer transmission efficiency, which also motivates this study. We thus define the problem of compressing the location data of a group of moving objects as the group data compression problem. Therefore, in this paper, we first introduce our distributed mining algorithm to approach the moving object clustering problem and discover group movement patterns. Then, based on the discovered group movement patterns, we propose a novel compression algorithm to tackle the group data compression problem. Our distributed mining algorithm comprises a Group Movement Pattern Mining (GMPMine) and a Cluster Ensembling (CE) algorithms. It avoids transmitting unnecessary and redundant data by transmitting only the local grouping results to a base station (the sink), instead of all of the moving objects' location data. Specifically, the GMP Mine algorithm discovers the local group movement patterns by using a novel similarity measure, while the CE algorithm combines the local grouping results to remove inconsistency and improve the grouping quality by using the information theory.

Mining

Mining May well have been the second of humankind's earliest endeavors—granted that agriculture was the first. The two industries ranked together as the primary or basic industries of early civilization. Little has changed in the importance of these industries since the beginning of civilization. If we consider fishing and lumbering as part of agriculture and oil and gas production as part of mining, then agriculture and mining continue to supply all the basic

Resources used by modern civilization. From prehistoric times to the present, mining has played an important part in human existence (Madigan,1981). Here the term *mining* is used in its broadest context as encompassing the extraction of any naturally occurring mineral substances—solid, liquid, and gas—from the earth or other heavenly bodies for utilitarian purposes. The most prominent of these uses for minerals are identified in Table 1.1. The history of mining is fascinating. It parallels the history of civilization, with many important cultural eras associated with and identified by various minerals or their derivatives: Many milestones in human history—Marco Polo's journey to China,Vasco da Gama's voyages to Africa and India, Columbus's discovery of the New World, and the modern gold rushes that led to the settlement of California, Alaska, South Africa, Australia, and the Canadian Klondike—were achieved with minerals providing a major incentive (Rickard,1932). Other interesting aspects of mining and metallurgical history can be found by referring to the historical record provided by Gregory (1980),Raymond (1984),and Lacy and Lacy (1992).

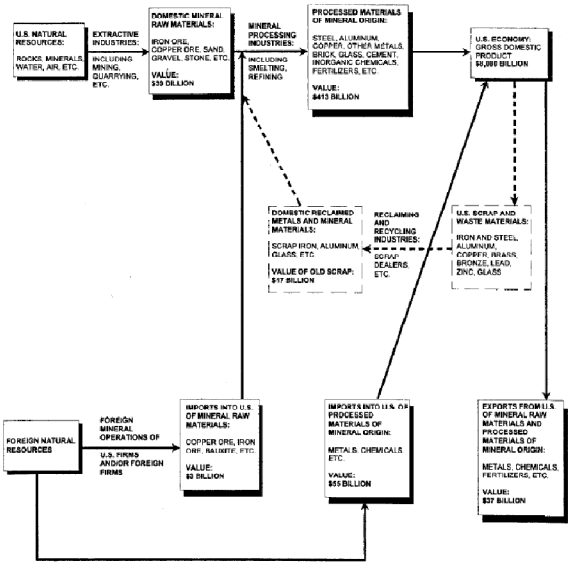

Fig 2.0 Block Diagram of mining

A ADVANCEMENTS IN MINING TECHNOLOGY

As one of humanity's earliest endeavors—and certainly one of its first organized industries—mining has an ancient and venerable history (Gregory, 1980). To understand modern mining practices, it is useful to trace the evolution of mining technology, which (as pointed out earlier in this chapter) has paralleled human evolution and the advance of civilization. Mining in its simplest form began with Paleolithic humans some 450,000 years ago, evidenced by the flint implements that have been found with the bones of early humans from the Old Stone Age (Lewis and Clark,1964). Our ancestors extracted pieces from loose masses of flint or from easily accessed outcrops and, using crude methods of chipping the flint, shaped them into tools and weapons. By the New Stone Age, humans had progressed to underground mining in systematic openings 2 to 3 ft (0.6 to 0.9m) in height and more than 30 ft (9m) in depth (Stoces,1954). However, the oldest known underground mine,a hematite mine at Bomvu Ridge, Swaziland (Gregory,1980),is from the Old Stone Age and believed to be about 40,000 years old. Early miners employed crude methods of ground

ISSN: 2277-9655
Impact Factor: 1.852

control,ventilation,haulage, hoisting, lighting, and rock breakage. Nonetheless,mines attained depths of 800 ft (250m) by early Egyptian times. Metallic minerals also attracted the attention of prehistoric humans. Initially, metals were used in their native form, probably obtained by washing river gravel in placer deposits. With the advent of the Bronze and Iron Ages, however, humans discovered smelting and learned to reduce ores into pure metals or alloys, which greatly improved their ability to use these metals. The first challenge for early miners was to break the ore and loosen it from the surrounding rock mass. Often, their crude tools made of bone, wood, and stone were no match for the harder rock, unless the rock contained crevices or cracks that could be opened by wedging or prying. As a result, they soon devised a revolutionary technique called *fire setting*, whereby they first heated the rock to expand it and then doused it with cold water to contract and break it. This was one of the first great advances in the science of rock breakage and had a greater impact than any other discovery until dynamite was invented by Alfred Nobel in 1867. Mining technology, like that of all industry, languished during the Dark Ages. Notably, a political development in 1185 improved the standing of mining and the status of miners, when the bishop of Trent granted a charter to miners in his domain. It gave miners legal as well as social rights, including the right to stake mineral claims. A milestone in the history of mining, the edict has had long-term consequences that persist to this day.

Data Compression

The typical sequence of operations performed in the compression of still images and video and audio data streams. The following example describes the compression of one image: The preparation step (her picture preparation) generates an appropriate digital representation of the information in the medium being compressed. For example, a picture might be divided into blocks of 8 pixels with a fixed number of bits per pixel. The processing step (here picture processing) is the first step that makes use of with various compression algorithms. For example, a transformation from the time domain to the frequency domain can be performed using the Discrete Cosine Transform (DCT). In the case of inter frame coding; motion vectors can be determined here for each 8-8 pixel block. Quantization takes place after the mathematically exact picture processing step. Values determined in the previous step cannot and should not be processed with full exactness; instead they are quantized according to a specific resolution and characteristic curve. This can also be considered equivalent to the law and *A*law, which are used for audio data [JN84]. In the transformed domain, the results can be treated differently depending on their importance (e.g., quantized with different numbers of bits). Entropy coding starts with a sequential data stream of individual bits and bytes.

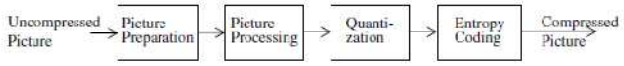

Fig 3.0 Block diagram of Data compression

Different techniques can be used here to perform a final, lossless compression. For example, frequently occurring long sequences of zeroes can be compressed by specifying the number of occurrences followed by the zero it. Picture processing and quantization can be repeated iteratively, such as in the case of Adaptive Differential Pulse Code Modulation (ADPCM). There can either be "feedback" (as occurs during delta modulation), or multiple techniques can be applied to the data one after the other (like interframe and intraframe coding in the case of MPEG). After these four compression steps, the digital data are placed in a data stream having defined format, which may also integrate the image starting point and type of compression. An error correction code can also be added at this point. Figure shows the compression process applied to a still image; the same principles can also be applied to video and audio data.

Coding Type	Basis	Technique
Entropy Coding	Run-length Coding	
	Huffman Coding	
	Arithmetic Coding	
Source Coding	Prediction	DPCM
		DM
	Transformation	FFT
		DCT
	Layered Coding (according to importance)	Bit Position
		Subsampling
		Subband Coding
	Vector Quantization	
Hybrid Coding	JPEG	
	MPEG	
	H.263	
	Many proprietary systems	

Table 1: Different Types of encoding techniques

Mining of Group Movement Pattern

To tackle the moving object clustering problem, we propose a distributed mining algorithm, which comprises the GMPMine and CE algorithms. First, the GMPMine algorithm uses a PST to generate an object's significant movement patterns and computes the similarity of two objects by using simp to derive the local

grouping results. The merits of simp include its accuracy and efficiency: First, simp considers the significances of each movement pattern regarding to individual objects so that it achieves better accuracy in similarity comparison. For a PST can be used to predict a pattern's occurrence probability, which is viewed as the significance of the pattern regarding the PST, sim thus includes movement patterns' predicted occurrence probabilities to provide fine-grained similarity comparison. Second, simp can offer seamless and efficient comparison for the applications with evolving and evolutionary similarity relationships. This is because simp can compare the similarity of two data streams only on the changed mature nodes of emission trees instead of all nodes. To combine multiple local grouping results into a consensus, the CE algorithm utilizes the Jaccard similarity coefficient to measure the similarity between a pair of objects, and normalized mutual information (NMI) to derive the final ensembling result. It trades off the grouping quality against the computation cost by adjusting a partition parameter. In contrast to approaches that perform clustering among the entire trajectories, the distributed algorithm discovers the group relationships in a distributed manner on sensor nodes. As a result, we can discover group movement patterns to compress the location data in the areas where objects have explicit group relationships. Besides, the distributed design provides flexibility to take partial local grouping results into ensembling when the group relationships of moving objects in a specified sub region are interested. Also, it is especially suitable for heterogeneous tracking configurations, which helps reduce the tracking cost, e.g., instead of waking up all sensors at the same frequency, a fine-grained tracking interval is specified for partial terrain in the migration season to reduce the energy consumption. Rather than deploying the sensors in the same density, they are only highly concentrated in areas of interest to reduce deployment costs.

A.The Group Movement Pattern Mining (GMPMine) Algorithm

To provide better discrimination accuracy, we propose a new similarity measure simp to compare the similarity of two objects. For each of their significant movement patterns, the new similarity measure considers not merely two probability distributions but also two weight factors, i.e., the significance of the pattern regarding to each PST. The similarity score simp of oi and oj based on their respective PSTs, Ti and Tj, is defined as follows:

$$sim_p(o_i, o_j) = -\log \frac{\sum_{s \in \tilde{S}} \sqrt{\sum_{\sigma \in \Sigma} (P^{T_i}(s\sigma) - P^{\bar{T}_j}(s\sigma))^2}}{2L_{max} + \sqrt{2}}, \quad (1)$$

where e S denotes the union of significant patterns (node strings) on the two trees. The similarity score simp includes the distance associated with a pattern s, defined as

$$d(s) = \sqrt{\sum_{\sigma \in \Sigma} (P^{T_i}(s\sigma) - P^{T_j}(s\sigma))^2}$$
$$= \sqrt{\sum_{\sigma \in \Sigma} (P^{T_i}(s) \times P^{T_i}(\sigma|s) - P^{T_j}(s) \times P^{T_j}(\sigma|s))^2},$$

The GMPMine algorithm includes four steps. First, we extract the movement patterns from the location sequences by learning a PST for each object. Second, our algorithm constructs an undirected, un weighted similarity graph where similar objects share an edge between each other. We model the density of group relationship by the connectivity of a sub graph, which is also defined as the minimal cut of the sub graph. When the ratio of the connectivity to the size of the sub graph is higher than a threshold, the objects corresponding to the sub graph are identified as a group.

Conclusion

In this work, we exploit the characteristics of group movements to discover the information about groups of moving objects in tracking applications. We propose a distributed mining algorithm, which consists of a local GMPMine algorithm and a CE algorithm, to discover group movement patterns. With the discovered information, we devise the 2P2D algorithm, which comprises a sequence merge phase and an entropy reduction phase. In the sequence merge phase, we propose the Merge algorithm to merge the location sequences of a group of moving objects with the goal of reducing the overall sequence length. In the entropy reduction phase, we formulate the HIR problem and propose ba Replace algorithm to tackle the HIR problem. In addition, we devise and prove three replacement rules, with which the Replace algorithm obtains the optimal solution of HIR efficiently. Our experimental results show that the proposed compression algorithm effectively reduces the amount of delivered data and enhances compressibility and, by extension, reduces the energy consumption expense for data transmission in WSNs.

References

[1] 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.

[2] W.-C. Peng and M.-S. Chen, "Developing Data Allocation Schemes by Incremental Mining of User Moving Patterns in aMobile Computing

System," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 1, pp. 70-85, Jan./Feb. 2003.
[3] M. Morzy, "Prediction of Moving Object Location Based on Frequent Trajectories," Proc. 21st Int'l Symp. Computer and Information Sciences, pp. 583-592, Nov. 2006.
[4] V. Guralnik and G. Karypis, "A Scalable Algorithm for Clustering Sequential Data," Proc. First IEEE Int'l Conf. Data Mining, pp. 179- 186, 2001.
[5] J. Yang and W. Wang, "CLUSEQ: Efficient and Effective Sequence Clustering," Proc. 19th Int'l Conf. Data Eng., pp. 101-112, Mar. 2003.
[6] J. Tang, B. Hao, and A. Sen, "Relay Node Placement in Large Scale Wireless Sensor Networks," J. Computer Comm., special issue on sensor networks, vol. 29, no. 4, pp. 490-501, 2006.
[7] M. Younis and K. Akkaya, "Strategies and Techniques for Node Placement in Wireless Sensor Networks: A Survey," Ad Hoc Networks, vol. 6, no. 4, pp. 621-655, 2008.
[8] S. Pandey, S. Dong, P. Agrawal, and K. Sivalingam, "A Hybrid Approach to Optimize Node Placements in Hierarchical Heterogeneous Networks," Proc. IEEE Conf. Wireless Comm. and Networking Conf., pp. 3918-3923, Mar. 2007.
[9] "Stargate: A Platform x Project," http://platformx.sourceforge. net, 2010.
[10] "Mica2 Sensor Board," http://www.xbow.com, 2010.
[11] J.N. Al-Karaki and A.E. Kamal, "Routing Techniques in Wireless Sensor Networks: A Survey," IEEE Wireless Comm., vol. 11, no. 6, pp. 6-28, Dec. 2004.