Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 26, S. 395-400, Temmuz 2021
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 26, pp. 395-400, July 2021
Copyright © 2021 EJOSAT
Research Article

BICOT: Big Data Analysis Approach for Clustering Cloud based IoT Systems

Zuleyha Akusta Dagdeviren[1]*, Orhan Dagdeviren[2]

[1*] Ege University, International Computer Institute, Izmir, Turkey (ORCID: 0000-0001-9365-326X), zuleyhaakusta@gmail.com
[2] Ege University, International Computer Institute, Izmir, Turkey (ORCID: 0000-0001-8789-5086), orhan.dagdeviren@ege.edu.tr

(3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications June 11-13, 2021)

(**DOI:** 10.31590/ejosat.960360)

ATIF/REFERENCE: Akusta Dagdeviren, Z. & Dagdeviren, O. (2021). BICOT: Big Data Analysis Approach for Clustering Cloud based IoT Systems. *European Journal of Science and Technology*, (26), 395-400.

Abstract

Internet of Things (IoT) envisions the connection of billions of devices over the Internet. The data produced by these huge amount of devices grow exponentially, so analyzing this big data with traditional methods is not viable. Recent cloud computing and virtualization technologies cope with these issues by processing and storing IoT data. Wireless sensor networks (WSNs) are big data sources of IoT systems which provides data collection from the environment. WSNs are used in various applications such as habitat monitoring, military surveillance and smart agriculture. Data transmission to the sink node is one of the essential requirements for WSNs. Clustering is a fundamental technique that is used for efficient data transmission, time synchronizaion, load balancing and security services. In this paper, we propose a clustering framework that we call BICOT for WSNs tailored for IoT systems. BICOT inputs large scale node position, transmission range and node energy data and outputs clustering information. Our first algorithm (BICOT-CDS) is based on connected dominating set (CDS) structure and aims to reduce the cluster count. Our second algorithm uses a weighted CDS (WCDS) approach that targets to select nodes with high energy as cluster heads. We implement these algorithms in ns2 simulator environment and measure cluster count and total weight of cluster head values. The algorithms are tested against node counts and average node degrees. From extensive simulation measurements, we obtain that the cluster count generated by BICOT-CDS is far more better than its counterparts and as the network size increases the proposed algorithm performs better. The cost of dominators produced by the BICOT-WCDS algorithm is significantly lower than its competitors. These findings show us that our proposed algorithms are favorable big data analysis approaches for cloud based IoT systems.

Keywords: Internet of Things, Big Data, Cloud Computing, Wireless Sensor Networks, Clustering, Dominating Set.

BICOT: Bulut Tabanlı IoT Sistemleri Kümeleme için Büyük Veri Analizi Yaklaşımı

Öz

Nesnelerin İnterneti (*Internet of Things:* IoT) milyarlarca cihazın İnternet üzerinden bağlanmasını öngörmektedir. Bu büyük miktardaki cihazların ürettiği veriler katlanarak büyümektedir, bu nedenle bu büyük veriyi geleneksel yöntemlerle analiz etmek mümkün olmamaktadır. Güncel bulut bilişim ve sanallaştırma teknolojileri, IoT verilerini işleyerek ve depolayarak bu sorunlarla başa çıkmaktadır. Kablosuz sensör ağlar (KSA'lar), ortamdan veri toplamayı sağlayan IoT sistemlerinin büyük veri kaynaklarıdır. KSA'lar, habitat izleme, askeri gözetim ve akıllı tarım gibi çeşitli uygulamalarda kullanılmaktadır. Çıkış düğümüne veri iletimi, KSA'lar için temel gereksinimlerden biridir. Kümeleme; verimli veri iletimi, zaman senkronizasyonu, yük dengeleme ve güvenlik servisleri için kullanılan temel bir tekniktir. Bu makalede IoT sistemleri için uyarlanmış KSA'lar için BICOT diye adlandırdığımız bir kümeleme çerçevesi önermekteyiz. BICOT, büyük ölçekli düğüm konumu, iletim alanı ve düğüm enerji verilerini girdi olarak almakta ve kümeleme bilgisini çıktı olarak üretmektedir. İlk algoritmamız (BICOT-CDS), bağlı hakim küme (*connected dominating set*: CDS) yapısına dayanmakta ve küme sayısını azaltmayı amaçlamaktadır. İkinci algoritmamız, küme başları olarak yüksek enerjiye sahip

[*] Sorumlu Yazar: zuleyhaakusta@gmail.com

düğümleri seçmeyi hedefleyen ağırlıklı bağlı hakim küme (*weighted connected dominating set:* WCDS) yaklaşımı kullanmaktadır. Bu algoritmaları ns2 simülatör ortamında uygulamakta ve küme sayısı ve küme başı değerlerinin toplam ağırlığını ölçmekteyiz. Algoritmalar, düğüm sayılarına ve ortalama düğüm derecelerine göre test ortamında benzetimleri yapılmaktadır. Kapsamlı simülasyon ölçümlerinden, BICOT-CDS tarafından üretilen küme sayılarının rakiplerinin ürettiği küme sayılarından çok daha iyi olduğunu ve ağ boyutu arttıkça önerilen algoritmanın daha iyi performans gösterdiğini elde etmekteyiz. BICOT-WCDS algoritması tarafından üretilen hakim düğümlerin maliyeti, rakiplerinin ürettiğinden önemli ölçüde daha düşüktür. Bu bulgular bize önerdiğimiz algoritmaların bulut tabanlı IoT sistemleri için uygun büyük veri analizi yaklaşımları olduğunu göstermektedir.

Anahtar Kelimeler: Nesnelerin İnterneti, Büyük Veri, Bulut Bilişim, Kablosuz Sensör Ağlar, Kümeleme, Hakim Küme.

1. Introduction

Internet of Things (IoT) is a network of billions of connected devices through the Internet infrastructure. Recent cloud computing and virtualization technologies provide processing and storage of big data produced by these vast amounts of IoT nodes. Wireless sensor networks (WSNs) are located at the communication layer of IoT technology and are assigned to collect various information from the environment. Thus, WSNs are important big data sources of IoT (Harb et al., 2017) (Kim et al., 2019). In sensor networks, clustering is a very important technique that provides efficient routing, data aggregation, time synchronization, load balancing, and security services (Lotfinezhad and Liang, 2005). Clustering is used to decrease the transmitted message count of the upper layer applications (Liu et al., 2020) (Vaiyapuri et al., 2021). In the clustering technique, nodes are either categorized as cluster members or cluster heads. Objectives for the clustering in WSNs are given as follows:

- A cluster head senses events from the environment, collects data from the cluster members and relays the aggregated data. Thus, it plays a server role of its cluster. Accordingly, it consumes much energy than the cluster members. In order to maintain a high network lifetime, cluster heads must be elected from nodes with high residual energy.
- Various physical layer and medium access control layer technologies exist to provide communication between distributed nodes. To provide independent execution, the clustering approach should not be decoupled tightly with an underlying networking standard.
- To increase the network lifetime by reducing resource consumption, the clustering algorithm should have low message and time complexities.
- Sensor nodes may have faults due to environmental effects, security attacks and hardware/software problems. In this manner, clustering operation should be fault tolerant to manage these external issues.

A WSN can be modeled with an undirected graph G=(V, E) as V and E are the set of vertices and edges, respectively. One of a techniques to provide clustering is construction of dominating sets (DSs). A DS S is a subset of V where a vertex $i \in S$ or ($j \in V$ and $(i,j) \in E$). In another words, a node is in DS or neighbor to a node in DS. If S induced subgraph is connected, we call S as connected dominating set (CDS). CDS is a very useful structure for routing in WSN since each dominator node is connected through each other. However, finding minimum CDS is an NP-hard problem. A CDS construction algorithm based on marking nodes due to some predefined rules is proposed in (Wu and Li, 1999). Guha and Khuller proposed a central CDS algorithm which provides growing of a tree T starting from vertex with the highest degree (Guha and Khuller, 1998). Initially, all nodes are marked as white. At each step, a pair of connected vertices is chosen to color black. Also, the colors of neighbors of these nodes are set to gray. The algorithm continues until a white node exists. Guha and Khuller's algorithm has an ln(n) approximation ratio for undirected graphs where n is the node count.

Although undirected graphs are useful structures for WSNs, they lack modelling some important parameters such as node energies. To overcome this problem, weighted undirected graph $G_w(V, E, w)$ can be used where w is a function to assign a weight value to each node. We represent the weights of the nodes as the reciprocal of their energies. Instead of minimizing the dominator count, the objective of minimum WCDS problem is minimizing the total weight to construct an energy-efficient routing infrastructure. A weighted maximal independent set (MIS) algorithm for construction of weighted DS is proposed in (Chatterjee et al., 2001). If a node has minimum weight ratio among its neighbors, it is selected as a dominator in Chatterjee et al.'s algorithm. In algorithm (Bao and Garcia-Luna-Aceves, 2003), a node enters DS if its weight ratio is minimum among its immediate neighbors or its weight ratio is minimum among its two-hop neighbors (immediate neighbors excluded). A weighted DS algorithm consisting of two phases is given in (Wang et al., 2006). The first phase starts with a MIS construction that resembles to (Chatterjee et al., 2001), then proceeds with a set cover algorithm (Chavatal, 1979). A minimum spanning tree algorithm is used in the second phase. Wang et al.'s algorithm is designed to execute in a distributed manner. Another central WCDS algorithm is given in (Guha and Khuller, 1998). Given M as the optimum solution, the approximation ratio of this algorithm is 3ln(M). Similar to the aforementioned algorithm, this algorithm consists of two stages. A set cover algorithm is used to construct weighted DS in the first stage (Chvatal, 1979), a Steiner tree algorithm is applied in the second stage (Klein and Ravi, 1995). For other studies related to clustering big data in IoT systems are given in (Palaniswami et al., 2020), (Tripathi et al., 2021) and (Wang et al., 2018).

In this paper, we propose clustering algorithms for IoT systems, namely BICOT. Our algorithms input large scale node position, transmission range and node energy data and output clustering information. These large scale data is stored in cloud where our big data analyzer system is located. Our proposed algorithms are CDS and WCDS based and they are executed in the cloud system. Our first algorithm (BICOT-CDS) aims to reduce the cluster count whereas the objective of our second algorithm (BICOT-WCDS) is to select cluster head nodes having higher energy than the other algorithms. We show the design of the algorithms and present measurements by comparing with their counterparts in the rest of this paper.

2. Material and Method

2.1. Network Model

Our network model is given in Fig. 1 by illustrating a sample IoT agriculture application.

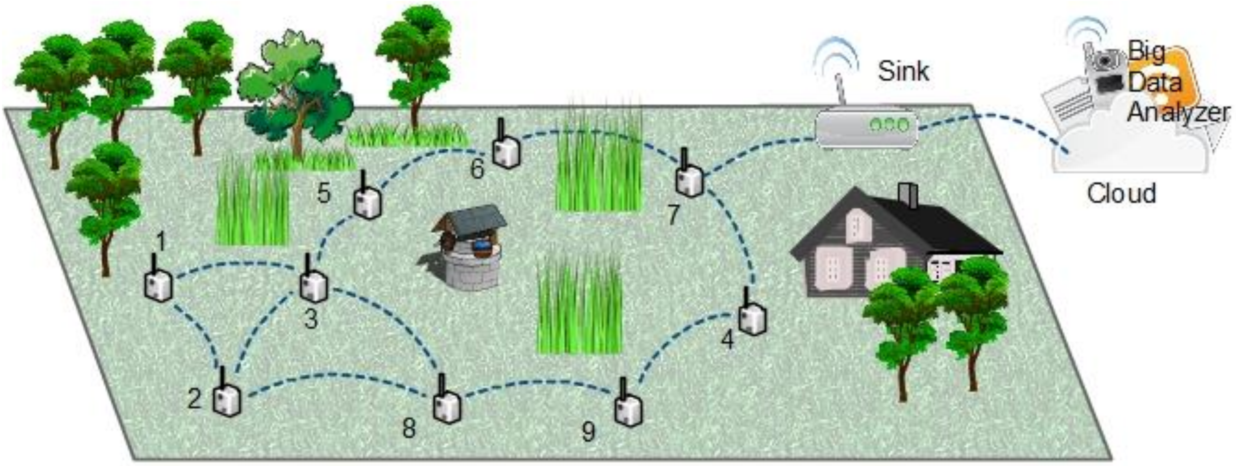


Figure 1. The Network Model

Please note that, the IoT vertical may vary such as smart city, habitat monitoring and military surveillance. On the other side, our network model is similar for other applications. To collect information from the environment such as humidity and mineral levels, we deploy a WSN in the field where nodes are given unique id.

In Fig. 1, there are 9 nodes in WSN and their communication channels are depicted with dotted arcs. Sensor nodes can be static or mobile, but they should be aware of their coordinates. This can be achieving by equipping a GPS receiver to each node or each node may execute localization algorithms to obtain positions. The coordinates of the sensor nodes should be collected by the sink node. A periodic flooding of positions may support this collection operation.

The sink node plays an IoT gateway role by connecting WSN and cloud system. Since a WSN may have thousands of mobile sensor nodes, frequent periodic collection of positions may cause to store a large scale data. So, this large scale data is stored in the cloud system and processed by a big data analyzer.

2.2. Proposed Algorithms

We propose two CDS algorithms for WSNs. The first algorithm aims to minimize the number of cluster heads, on the other hand the aim of the second algorithm is to accomplish energy efficient cluster head selection. The steps of the first algorithm, which we call as BICOT-CDS is given in Alg. 1.

Algorithm 1. BICOT-CDS Algorithm

```
input: coordinates of nodes and the transmission range ($T$).
construct graph $G=(V, E)$ using transmission range and coordinates of nodes (If $(x_i-x_j)^2+(y_i-y_j)^2 \leq T^2$ then $node_i$ and $node_j$ are neighbors in $G$).
call a central connected dominating set algorithm and output the result to $S$.
connect every cluster member to the closest cluster head in $S$. If there are more than one, choose the node with the greatest id.
output: $S$ is the backbone.
```

We first construct an undirected graph and give this graph as an input to a central CDS algorithm. At the last step of BICOT-CDS algorithm, each cluster member is connected to a cluster head. An example WSN clustered with BICOT-CDS is given in Fig. 2 (The algorithm in (Guha and Khuller, 1998) is used as the central CDS algorithm). In this figure, cluster heads are shown with red nodes where each of them is an element of CDS and the white nodes are ordinary cluster members. In this manner, nodes C, E, I, J, M, P and R are dominators and other nodes are dominatees. The clusters are depicted with dotted orange circular drawings.

Algorithm 2. BICOT-WCDS Algorithm

```
input: coordinates of the nodes, transmission range ($T$) and the energies of nodes.
construct node weighted graph $G_w=(V, E)$ using transmission range, coordinates of nodes (If $(x_i-x_j)^2+(y_i-y_j)^2 \leq T^2$ then $node_i$ and $node_j$ are neighbors in $G$) and energies of the nodes (weight($node_i$) = 1 /energy($node_i$)).
call a central weighted connected dominating set algorithm and output the result to $S_w$.
connect every cluster member to the closest cluster head in $S_w$ with the maximum energy as a cluster member. If there are more than one, choose the nearest cluster head node.
output: $S_w$ is the weighted backbone.
```

The second algorithm resembles the first algorithm whereas the dominating set is constructed in an energy efficient manner. The steps of this algorithm, which we call as BICOT-WCDS, are given in Alg. 2. In this algorithm we use a central weighted algorithm such as (Guha and Khuller, 1998) to produce the cluster heads. Those cluster heads are the elements of the weighted backbone.

Fig. 3 displays a sample operation of the proposed algorithm (The central WCDS algorithm in (Guha and Khuller, 1998) is used). In this figure, node ids and weights are written near to the nodes. Similar, to the previous figure the borders of the clusters are depicted with orange circular drawings. The WCDS backbone consists of nodes A, C, D, I, J, K, M, O and S. As seen in this figures, nodes having low weights are chosen as the cluster heads. Nodes F, G, L, B, E, H, N, P are the cluster members of these cluster heads. Each cluster member node is connected to a cluster head having higher energy.

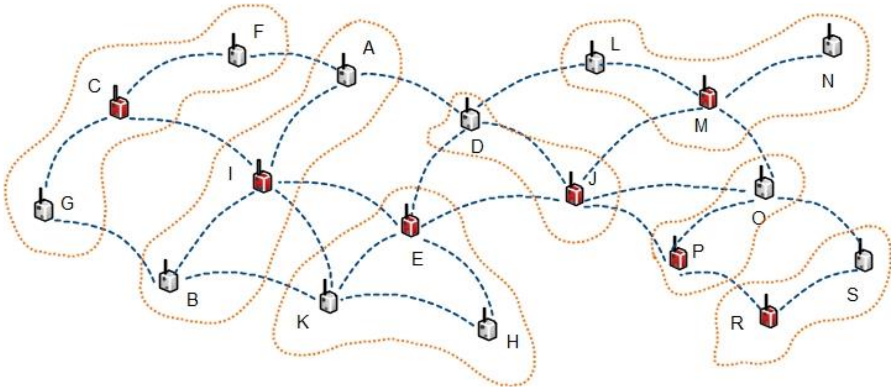


Figure 2. An Example WSN Clustered with BICOT-CDS

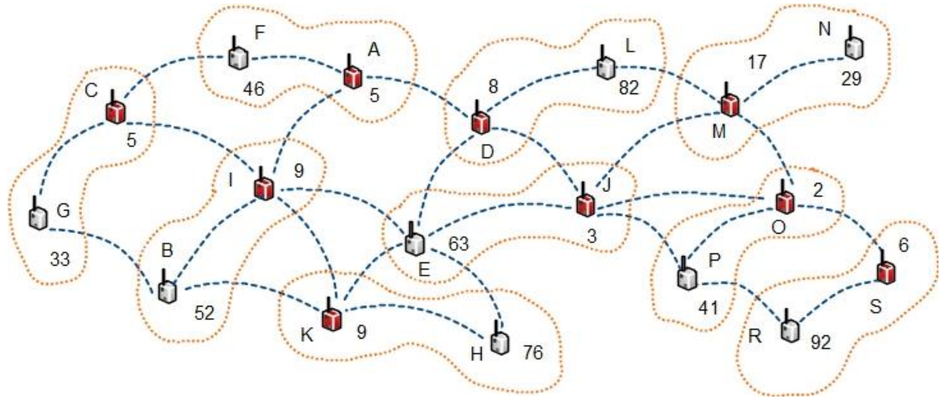


Figure 3. An Example WSN Clustered with BICOT-WCDS

3. Results and Discussion

We implement our proposed BICOT-CDS and BICOT-WCDS algorithms in ns2 simulator (VINT, 2021). The algorithms tested against various node counts and degrees. Node counts are varied from 100 nodes to 400 nodes. Average node degrees of the networks are selected from 4 to 6. IEEE 802.11 physical layer and medium access control layers are used as the underlying networking protocols. We implement a distributed CDS algorithm (DIST-CDS) to compare with the BICOT-CDS. DIST-CDS is similar with the algorithm (Wu and Li, 1999). DIST-CDS has two rules and each node is white at the beginning. In the first rule, a node *v* is colored black, node *v* has a pair of unconnected neighbors. In the second rule, a node *v* is colored white, if it has neighbors with greater id that span its 1-hop neighbors. After the execution of the rules, black nodes constitute the dominating set. MIS-WCDS (Chatterjee et al., 2001), BAO-WCDS (Bao and Garcia-Luna-Aceves, 2003) and WANG-WCDS (Wang et al., 2006) are implemented to compare with our BICOT-WCDS algorithm.

To measure the clustering quality of the CDS algorithms, we use the number of clusters as the cluster quality metric since CDS algorithms are generally targeted to reduce the set size to approximate minimum CDS. When degree is increased, the cluster count produced by BICOT-CDS is decreased as shown in Fig. 4.a since a cluster head may dominate more nodes in dense networks. As seen in Fig. 4.b, the cluster count generated by BICOT-CDS is approximately half of the number of clusters generated by DIST-CDS. As the network size increases, BICOT-CDS performs surely better than DIST-CDS.

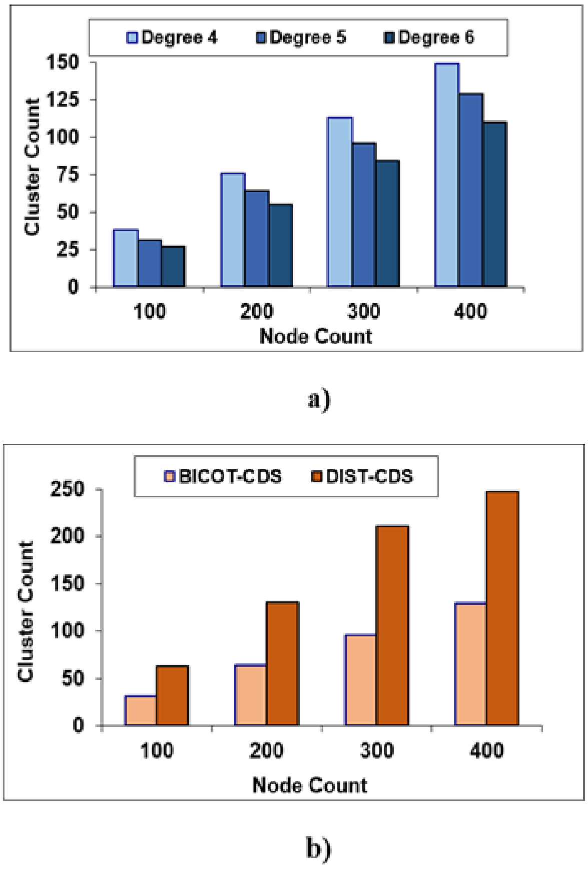
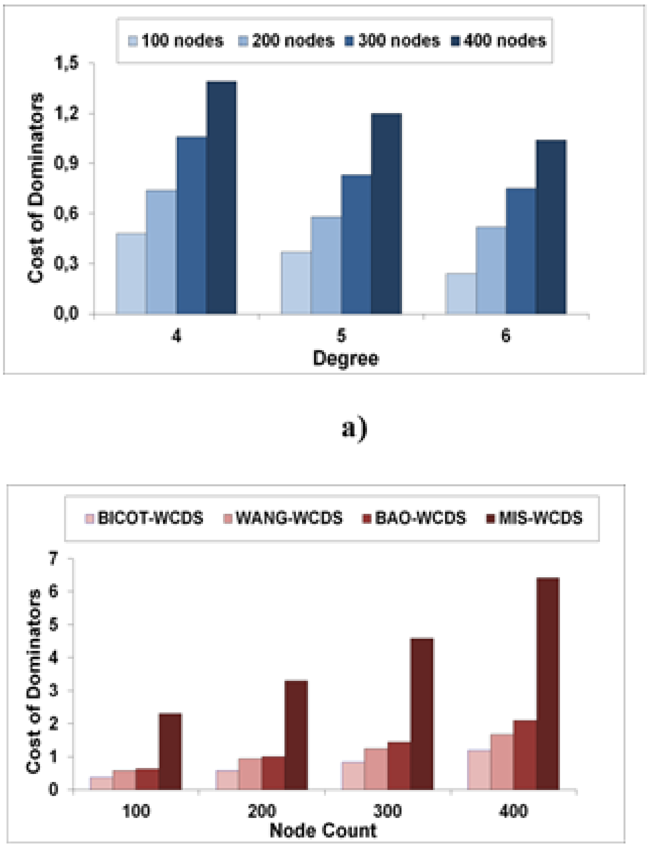


Figure 4. a) Cluster Count of BICOT-CDS vs. Node Count and Degree b) Cluster Count of Algorithms vs. Node Count

As aforementioned, we use Guha and Khuller's WDCS algorithm in our BICOT-WCDS algorithm. In the first phase of this algorithm, dominators are produced. In the second phase of the algorithm, connectors are provided. In Fig. 5.a, the dominator weights of the BICOT-WCDS vs. varying node degrees are given. As the average degree is rised, the proposed algorithm outputs dominators having higher energy. Since the cost is mapped to the reciprocal of node energy, the total cost of selected dominators decreases. This reveals that BICOT-WCDS reacts good with the rise of connectivity. Also, as the number of nodes is rised, the cumulative weight increases linearly as given in Fig. 5.a. In Fig. 5.b, cost of dominators produced by the algorithms against varying node count and degree are given. BICOT-WCDS has the best performance among other WCDS algorithms in all cases.

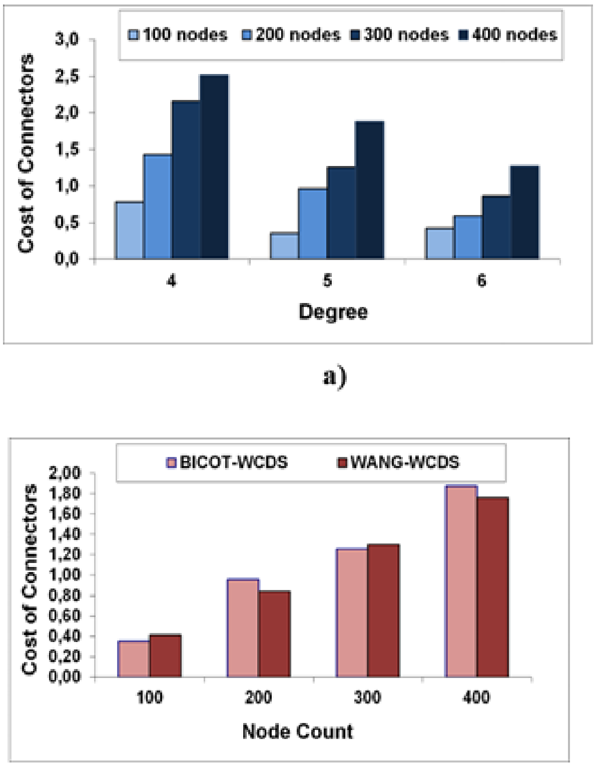


a)

b)

Figure 5. a) Cost of Dominators of BICOT-WCDS vs. Node Count and Degree b) Cost of Dominators of Algorithms vs. Node Count

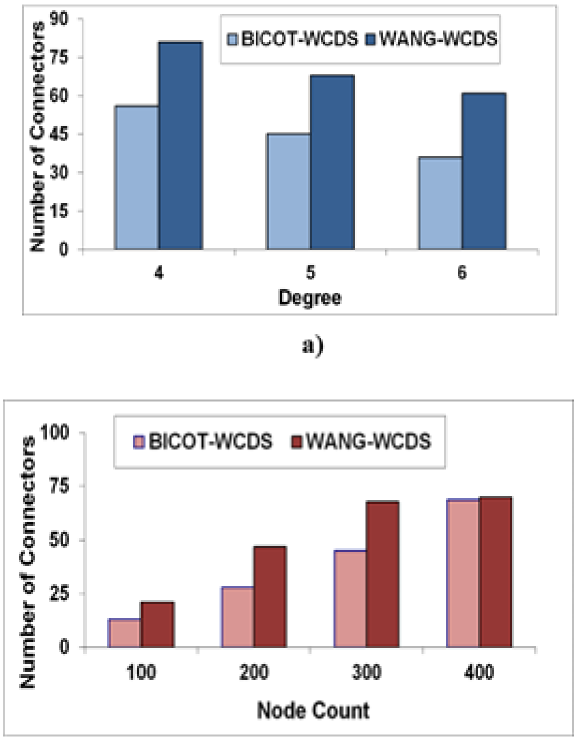
The cost of connectors of BICOT-WCDS vs. varying node degree values are given in Fig. 6.a. The cumulative connector weight decreases as the average node degree rises, as expected. Total connector weight of algorithms vs. varying node counts are shown in Fig. 6.b. For node counts equal to 100 and 300 BICOT-WCDS performs better whereas for node counts equal to 200 and 400 the performance of WANG-WCDS is better. Consequently, the performance of the algorithms are generally approximate. The performance comparison of the BICOT-WCDS and WANG-WCDS are given in Fig. 7.a and Fig. 7.b, where the connector counts produced by BICOT-WCDS are smaller than those of WANG-WCDS vs. node count and average node degree.



a)

b)

Figure 6. a) Cost of Connectors of BICOT-WCDS vs. Node Count and Degree b) Cost of Connectors of Algorithms vs. Node Count



a)

b)

Figure 7. a) Number of Connectors of Algorithms vs. Degree b) Number of Connectors of Algorithms vs. Node Count

4. Conclusions and Recommendations

IoT brings the opportunity of connection of billions of devices over the Internet. WSNs are big data resources of IoT systems where cloud technologies handle the data produced by the sensor nodes. Clustering is a very important method used for routing, load balancing and time synchronization in WSNs. In this paper, we propose BICOT clustering framework for WSNs that is based on CDS structure. Our framework consists of two algorithms: BICOT-CDS and BICOT-WCDS. BICOT-CDS aims to reduce the cluster count whereas BICOT-WCDS targets to select nodes with high energy as cluster heads.

We implement BICOT-CDS and BICOT-WCDS algorithms in ns2 simulator environment against varying node counts and average node degrees. We measure cluster counts and total weight of cluster heads. From extensive measurements, we reveal that the cluster counts produced by BICOT-CDS algorithm is significantly better than its competitors and as the network size increases the performance gap of the algorithms increases. The cumulative cost of dominators produced by BICOT-WCDS algorithm is lower than those of its counterparts. Consequently, these findings reveal us that the proposed algorithm is a promising technique for clustering large scale position data for cloud based IoT systems.

References

Bao, L. and Garcia-Luna-Aceves, J. J. (2003) Topology management in ad hoc networks. *Proc. of the 4th ACM Int. Symp. on Mobile Ad Hoc Networking & Computing*, pp. 129-140, ACM Press, New York.

Chatterjee, M., Das, S. K., and Turgut, D. (2001) WCA: weighted clustering algorithm for mobile ad hoc networks. *Journal of Cluster Computing (Special Issue on Mobile Ad hoc Networks)*, 5, 193-204.

Chvatal, V. (1979) A greedy heuristic for the set-covering problem, Mathematics of Operations Research. *INFORMS*, 4(3), 233-235.

Guha, S. and Khuller, S. (1998) Approximation algorithms for connected dominating sets. *Algorithmica*, 20, 374-387.

Harb, H., Makhoul, A., Idrees, A., Zahwe and O. and Taam, M.. (2017) Wireless Sensor Networks: A Big Data Source in Internet of Things. *International Journal of Sensors, Wireless Communications and Control*.

Kim, B.-.S, Kim, K.-I., Shah, B., Chow, F. and Kim, K. H. (2019) Wireless Sensor Networks for Big Data Systems, *Sensors* 19, no. 7, 1565.

Klein, P. and Ravi, R. (1995) A nearly best-possible approximation algorithm for node-weighted steiner trees. *J. Algorithms*, 19(1), 104-105.

Liu, X., Zhu, R., Anjum, A., Wang, J., Zhang, H. and Ma, M. (2020) Intelligent data fusion algorithm based on hybrid delay-aware adaptive clustering in wireless sensor networks, *Future Generation Computer Systems*, vol.104, pp. 1-14,

Lotfinezhad, M. and Liang, B. (2005) Energy efficient clustering in sensor networks with mobile agents. *Proc. of the IEEE Wireless Communications and Networking Conf.*, New Orleans, USA, 13-17 March, pp. 1872-1877. IEEE, Washington.

Palaniswami, M., Rao, A. S., Kumar, D., Rathore, P. and Rajasegarar, S., (2020) The Role of Visual Assessment of Clusters for Big Data Analysis: From Real-World Internet of Things, *IEEE Systems, Man, and Cybernetics Magazine*, vol. 6, no. 4, pp. 45-53.

Tripathi, A. K., Sharma, K., Bala, M., Kumar, A., Menon, V. G. and Bashir, A. K. (2021) A Parallel Military-Dog-Based Algorithm for Clustering Big Data in Cognitive Industrial Internet of Things, *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2134-2142.

Wang, Q., Guo, S., Hu, J. and Yang, Y., (2018) Spectral partitioning and fuzzy C-means based clustering algorithm for big data wireless sensor networks. *EURASIP Journal on Wireless Communications and Networking*, 54.

Wang, Y., Wang, W., and Li, X.-Y. (2006) Efficient distributed low-cost backbone formation for wireless networks. *IEEE Trans. on Parallel and Dist. Syst.*, 17(7), 681-693.

Wu, J. and Li, H. (1999) On calculating connected dominating set for efficient routing in ad hoc wireless networks. *Proc. of the 3rd Int. Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications*, Seattle, Washington, United States, pp. 7-14. ACM, New York.

Vaiyapuri, T., Parvathy, V.S., Manikandan, Krishnaraj, V. N., Gupta, D. and Shankar, K. (2021) A Novel Hybrid Optimization for Cluster-Based Routing Protocol in Information-Centric Wireless Sensor Networks for IoT Based Mobile Edge Computing. Wireless Personal Communications, https://doi.org/10.1007/s11277-021-08088-w.

VINT project. (2021) Network Simulator version 2 (NS-2). *Technical Report*, available from: http://nsnam.sourceforge.net/wiki/ index.php/Main_Page.