



UNIVERSITY OF
THESSALY



Intelligent Pervasive Systems Research Group
Dept. of Informatics and Telecommunications

Query Driven Data Subspace Mapping

Panagiotis Fountas, Maria Papathanasaki, Kostas
Kolomvatsos, Christos Anagnostopoulos
Email: {[pfountas](mailto:pfountas@uth.gr), [mpapathanasaki](mailto:mpapathanasaki@uth.gr), [kostasks](mailto:kostasks@uth.gr)}@uth.gr,
christos.anagnostopoulos@glasgow.ac.uk

AIAI 2022

17 – 20 June 2022, Crete, Greece

18th International Conference on Artificial Intelligence Applications and Innovations.

Huge Volumes of Data



A tremendous number of Internet of Things (IoT) devices, computers, and applications produce data at a humongous rate and in large quantities.

Data Mapping

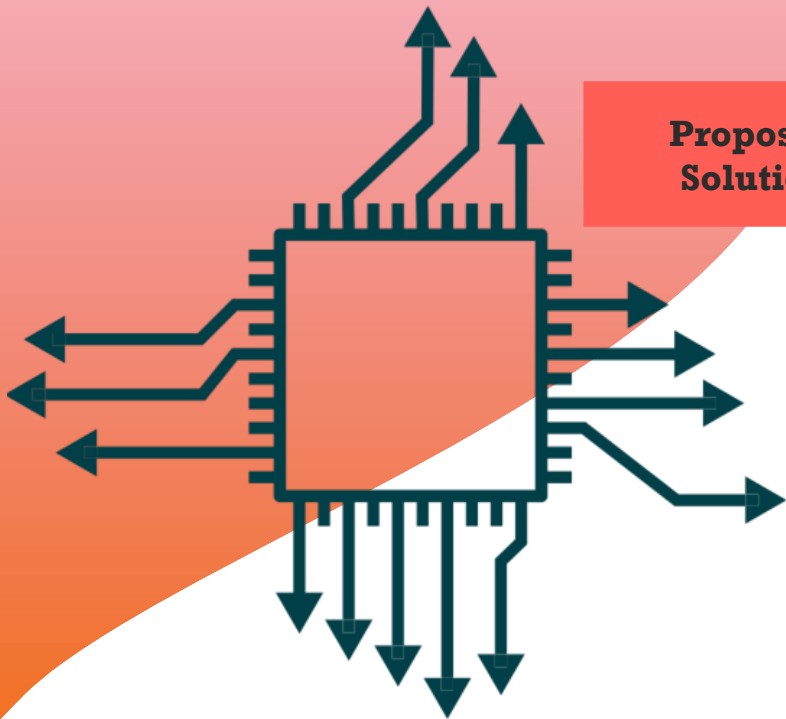


The process of data mapping is very important due to the storage of data in geo-distributed datasets and the need for fast and accurate responses in queries

Proposed Solution



We propose a hierarchical clustering model which detects the required data to respond in queries in the minimum possible time



Problem Description



IoT devices/applications collect multivariate data.



Distributed datasets host the data that produced by the IoT devices.



The servers receive queries from users and applications

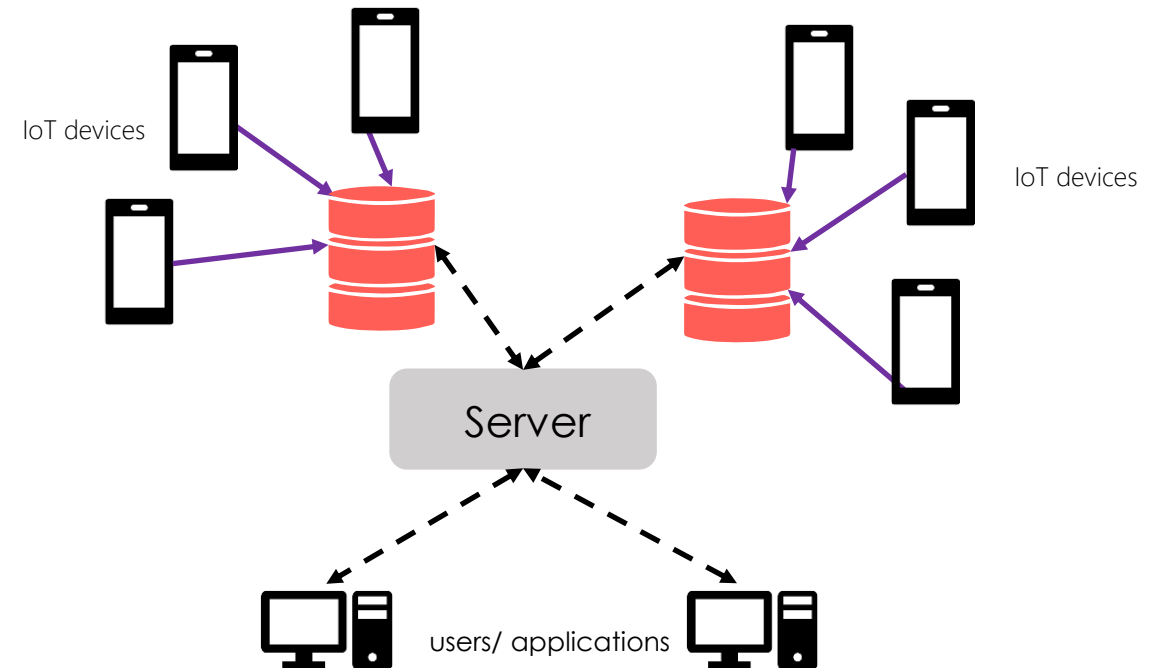


The servers return the appropriate data to the users/applications



Servers use the proposed model to map the required data for the incoming queries efficiently and in short time.

Time instance	1 st dimension	2 nd dimension	...	d th dimension
0	$x_1^j[0]$	$x_2^j[0]$...	$x_d^j[0]$
1	$x_1^j[1]$	$x_2^j[1]$...	$x_d^j[1]$
...
W	$x_1^j[W]$	$x_2^j[W]$...	$x_d^j[W]$



- ⚙️ Two phases: (i) the 'warm up' phase; (ii) the 'performance' phase.
- ⚙️ The 'warm up' phase consists of two stages: (i) the preparation stage; (ii) the hierarchical clustering execution stage.
- ⚙️ In the preparation stage, the HMCM performs a sequential scan of the entire database to determine the suitable data points.
- ⚙️ When a z number of queries have been sent to the server, the HMCM proceeds to the hierarchical clustering execution stage to create clusters and subclusters upon historical queries.
- ⚙️ HMCM adopts the Fuzzy C-Means to create a set of clusters \mathcal{C} upon z queries. Afterwards, it divides every cluster further into a set of $N_{C_i} = \{S_1, S_2, \dots, S_M\}$ using the K-Means algorithm.



Hierarchical Mixed Clustering Model (HMCM)

- ⚙️ In the 'performance' phase, the HMCM is activated every time a query is reported to the server.
- ⚙️ The HMCM uses Euclidean Distance to identify the top- r clusters whose members have the same data requests as the incoming query.
- ⚙️ The HMCM utilizes again Euclidean Distance to detect, for each of the top- r clusters, the top- ℓ subclusters.
- ⚙️ We focus only on groups of queries that have the most relevant data requests to the incoming query.

- ⚙️ We propose the adoption of an overlapping metric to detect the ‘matching’ between the data requests of queries. The Area Overlap Metric (AOM) is provided by the following equation:




$$\text{AOM}(q_{inc}, q_{member}) = \frac{q_{inc} \cap q_{member}}{\text{incoming query area}}$$



- ⚙️ The HMCM examines the members of the detected subclusters, to find those queries q_{member} with which the AOM overcomes a threshold θ and retrieves only the data points that belong to them.






Baseline Method (BM):

-  It is executed every time that a query is coming to the server.
-  It scans all the data sequentially to detect those that satisfy the query.
-  It is the theoretical optimal threshold in terms of error.



Hard Clustering Based Method (HCBM):

-  It is trained over z incoming queries (training phase), using the BM to identify the data required for their execution, and then, utilizes K-Means to cluster them.
-  It is executed every time that a query is coming to the server.
-  It uses Euclidean distance to find the top- r similar clusters to the incoming query and retrieve the data points that satisfy it.

Dataset	Source
Query Analytics Workloads Dataset	http://archive.ics.uci.edu/ml/datasets/Query+Analytics+Workloads+Dataset

- ⚙️ Warming dataset D_w : 1.000 range queries of the format $q_i = \{X_i, Y_i, Xr_i, Yr_i\}$.
- ⚙️ Dataset of two-dimensional points (D_{2d}) : 24.923 random generated 2d points
- ⚙️ Test dataset (D_T): $\psi=1.000$ incoming Range queries with the same distribution and format with the D_w

True Positive (TP): the number of data that an incoming query q_i needs and they detected

True Negative (TN): the number of data that the q_i does not need and the model correctly reject them

False Positive (FP): the number of data that the q_i does not need but the model retrieved them

False Negative (FN): the number of data that the q_i demands but the model does not detect them

Performance metrics:

$$\text{Precision (PRE)} = \frac{TP}{TP+FP}$$

$$\text{Recall (REC)} = \frac{TP}{TP+FN}$$

$$\text{Accuracy (ACC)} = \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN}$$

$$F_1 \text{ score (FSC)} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

Time per query (τ) = time for the mapping of data per query

The final performance is defined as follows:

$$\mu_{\Omega} = \frac{\sum_{q_i \in D_T} \Omega_i}{\psi}, \Omega \in \{\text{PRE, REC, ACC, FPR, FSC, } \tau\}$$

Γ : the number of clusters which are created in HMCM using Fuzzy C-Means

M : the number of subclusters which are created in HMCM using K-Means

K : the number of clusters which are created in HCBM using K-Means

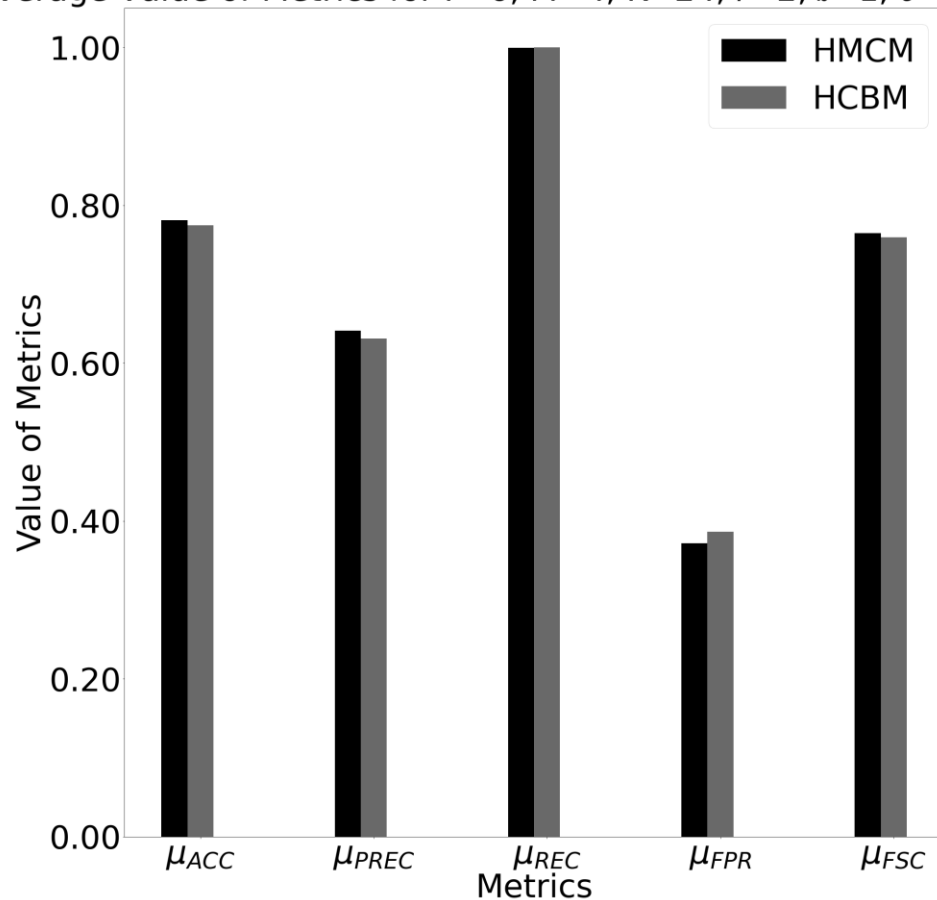
r : the number of top similar clusters

ℓ : the number of top similar subclusters

θ : the overlap threshold

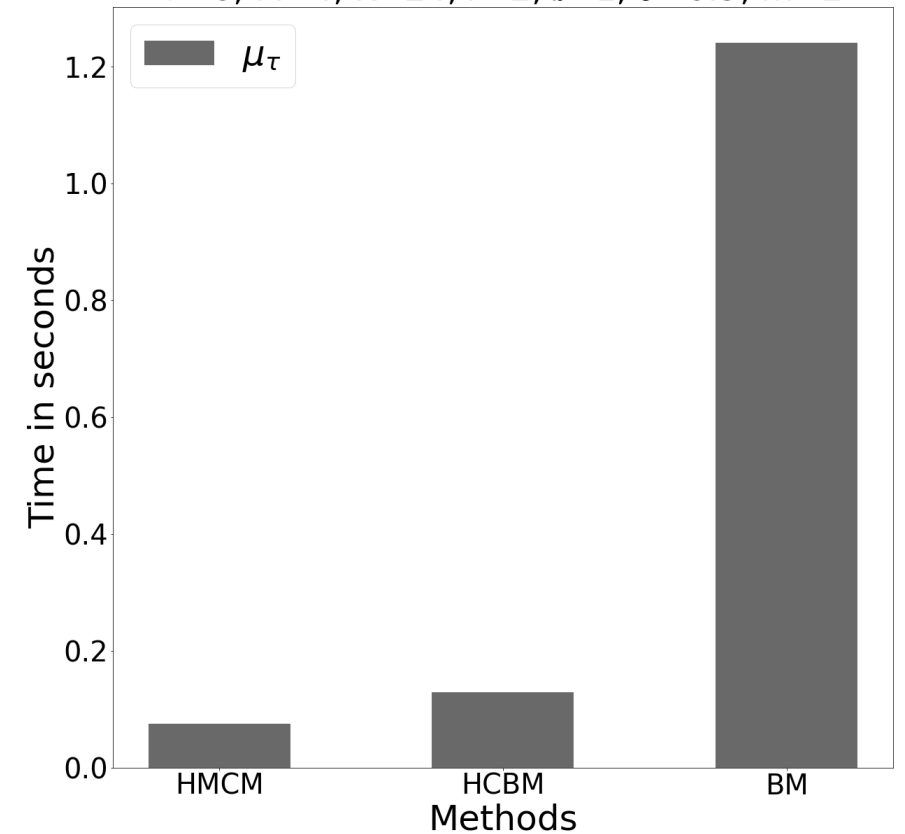
m : the fuzziness in Fuzzy C-Means

Average Value of Metrics for $\Gamma=6, M=4, K=24, r=2, \ell=1, \theta=0.9, m=2$



μ_{τ} for the mapping of data for each query

$\Gamma=6, M=4, K=24, r=2, \ell=1, \theta=0.9, m=2$



Γ : the number of clusters which are created in HMCM using Fuzzy C-Means

M : the number of subclusters which are created in HMCM using K-Means

K : the number of clusters which are created in HCBM using K-Means

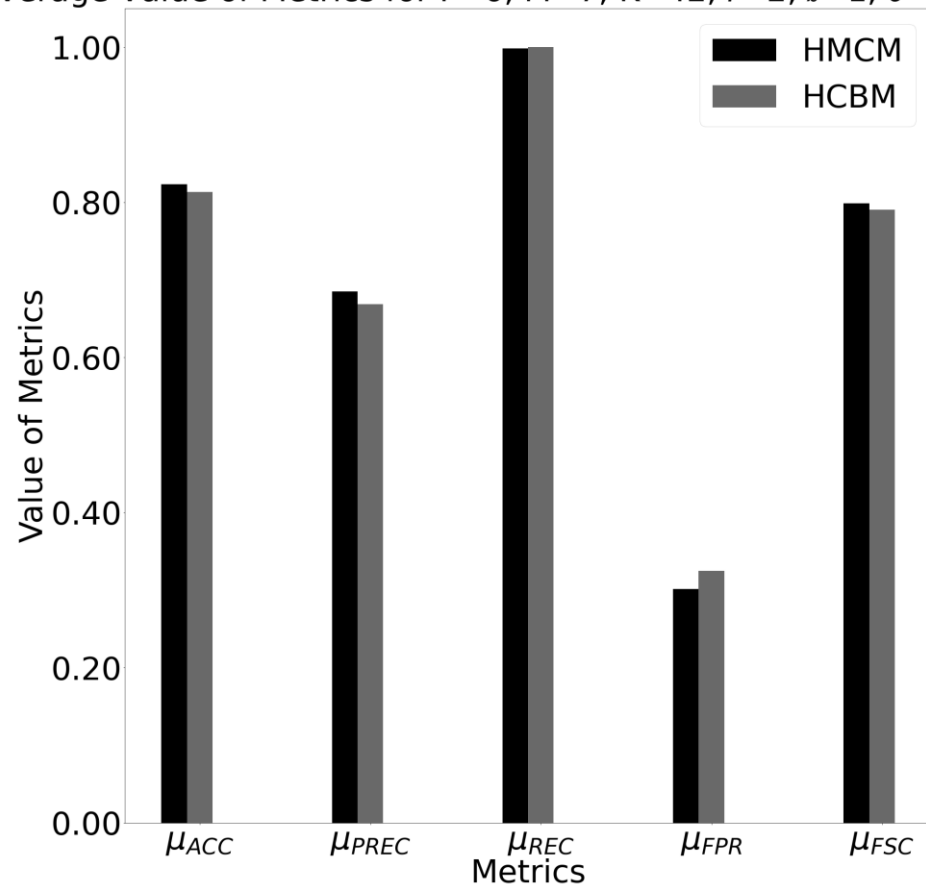
r : the number of top similar clusters

ℓ : the number of top similar subclusters

θ : the overlap threshold

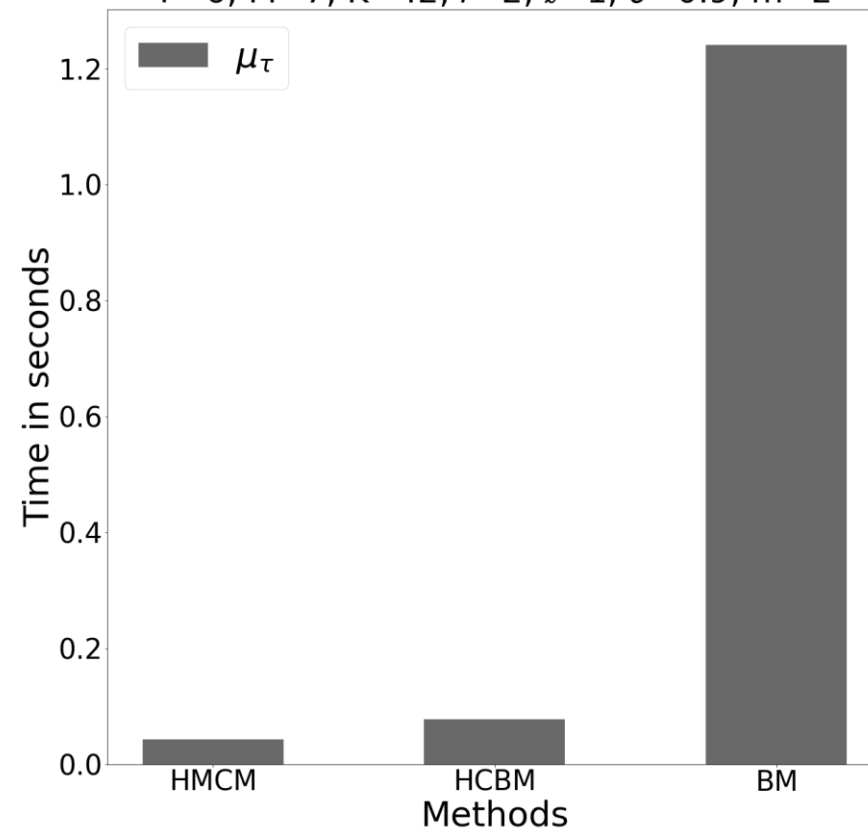
m : the fuzziness in Fuzzy C-Means

Average Value of Metrics for $\Gamma=6, M=7, K=42, r=2, \ell=1, \theta=0.9, m=2$



μ_{τ} for the mapping of data for each query

$\Gamma=6, M=7, K=42, r=2, \ell=1, \theta=0.9, m=2$



Γ : the number of clusters which are created in HMCM using Fuzzy C-Means

M : the number of subclusters which are created in HMCM using K-Means

K : the number of clusters which are created in HCBM using K-Means

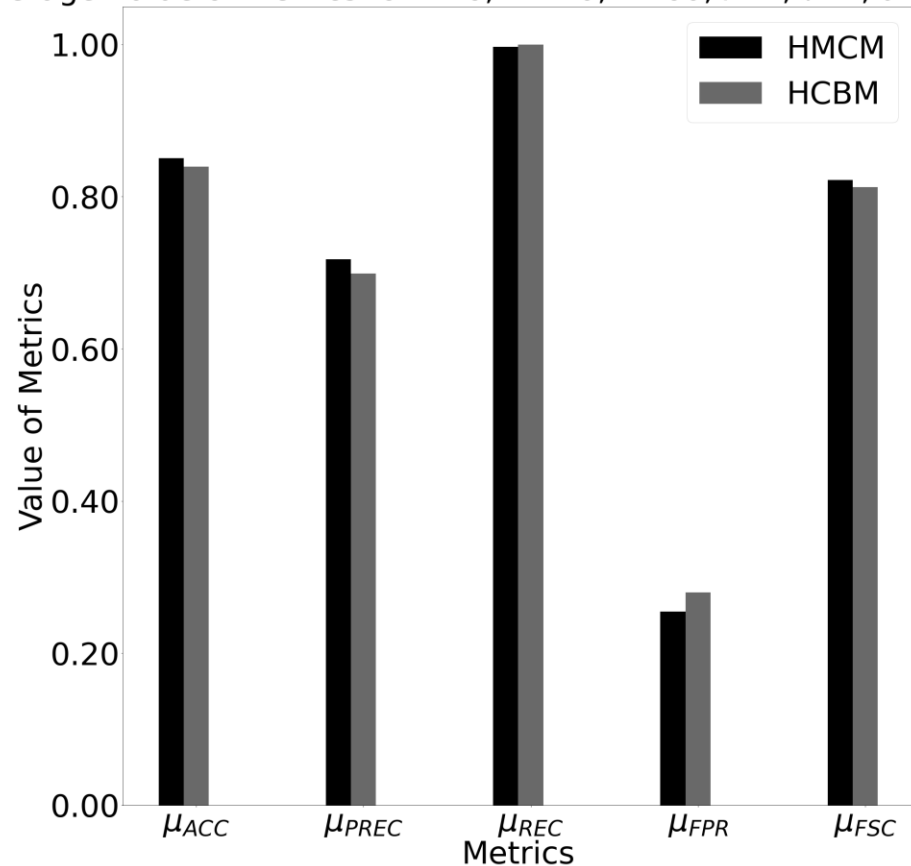
r : the number of top similar clusters

ℓ : the number of top similar subclusters

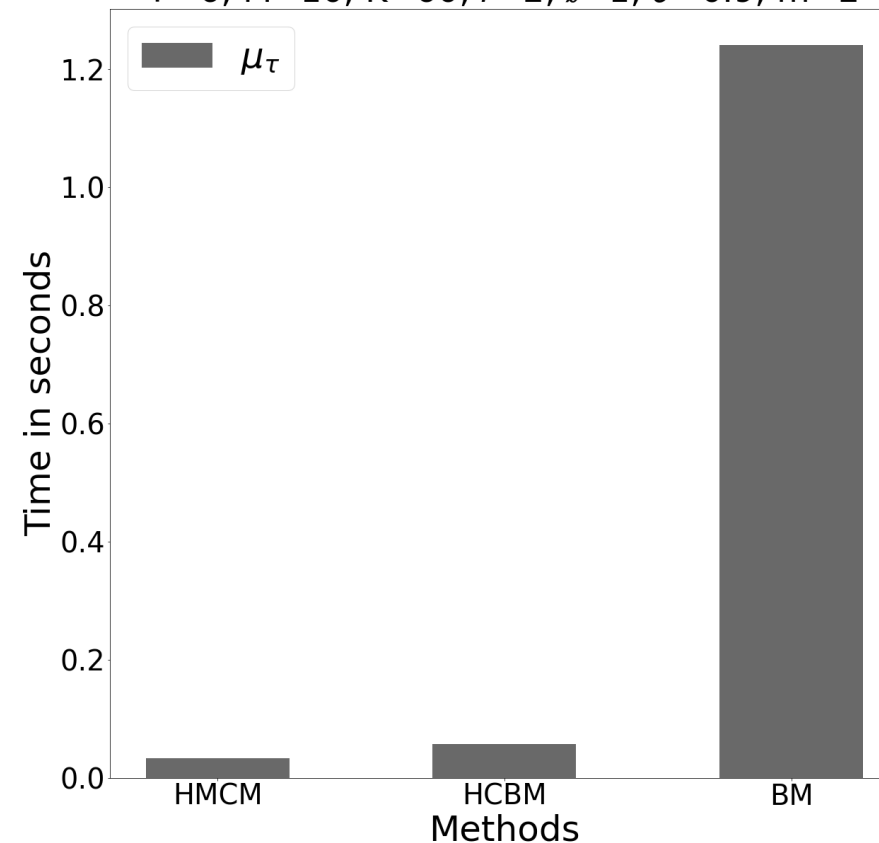
θ : the overlap threshold

m : the fuzziness in Fuzzy C-Means

Average Value of Metrics for $\Gamma=6, M=10, K=60, r=2, \ell=1, \theta=0.9, m=2$



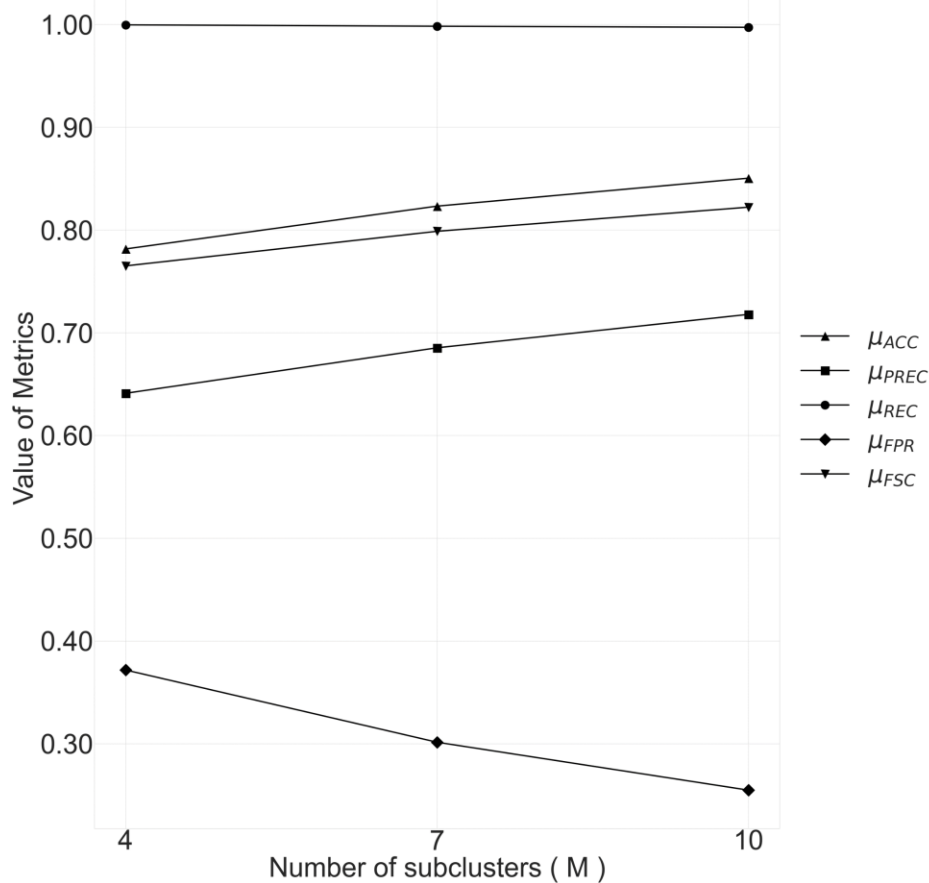
μ_{τ} for the mapping of data for each query
 $\Gamma=6, M=10, K=60, r=2, \ell=1, \theta=0.9, m=2$



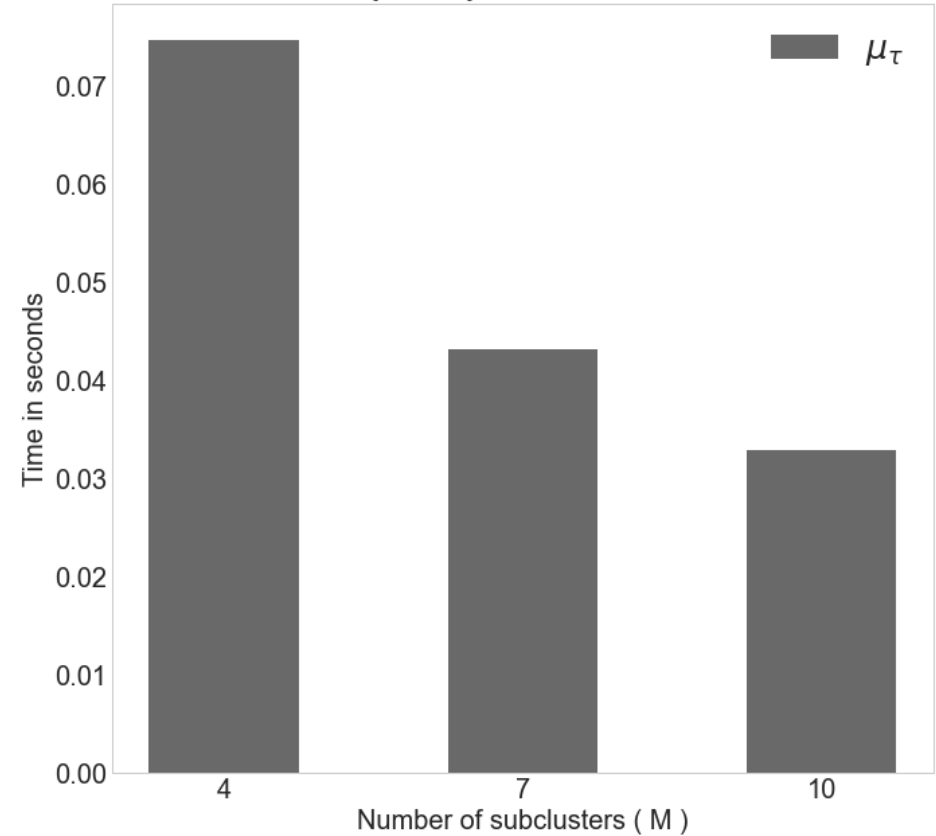
Γ : the number of clusters which are created in HMCM using Fuzzy C-Means
 M : the number of subclusters which are created in HMCM using K-Means
 K : the number of clusters which are created in HCBM using K-Means

r : the number of top similar clusters
 ℓ : the number of top similar subclusters
 θ : the overlap threshold
 m : the fuzziness in Fuzzy C-Means

Average Value of Metrics for $\Gamma=6$, $M=\{4, 7, 10\}$, $r=2$, $\ell=1$, $\theta=0.9$, $m=2$



μ_{τ} for the mapping of data for each query
 $\Gamma=6$, $M=\{4,7,10\}$, $r=2$, $\ell=1$, $\theta=0.9$, $m=2$



Γ : the number of clusters which are created in HMCM using Fuzzy C-Means

M : the number of subclusters which are created in HMCM using K-Means

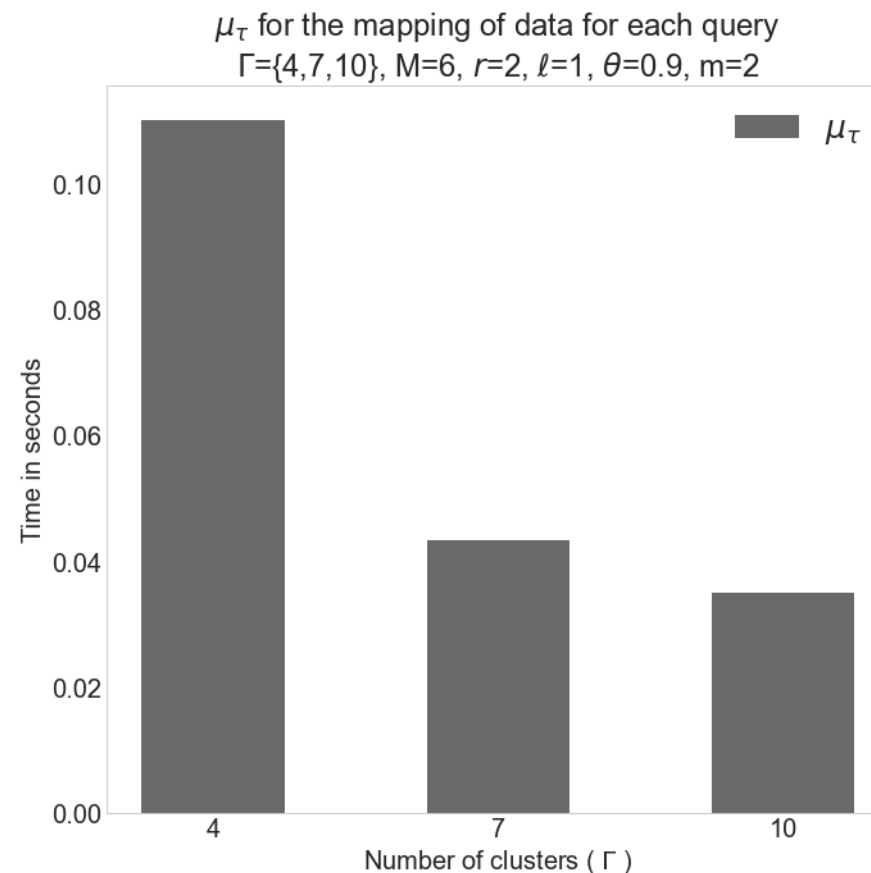
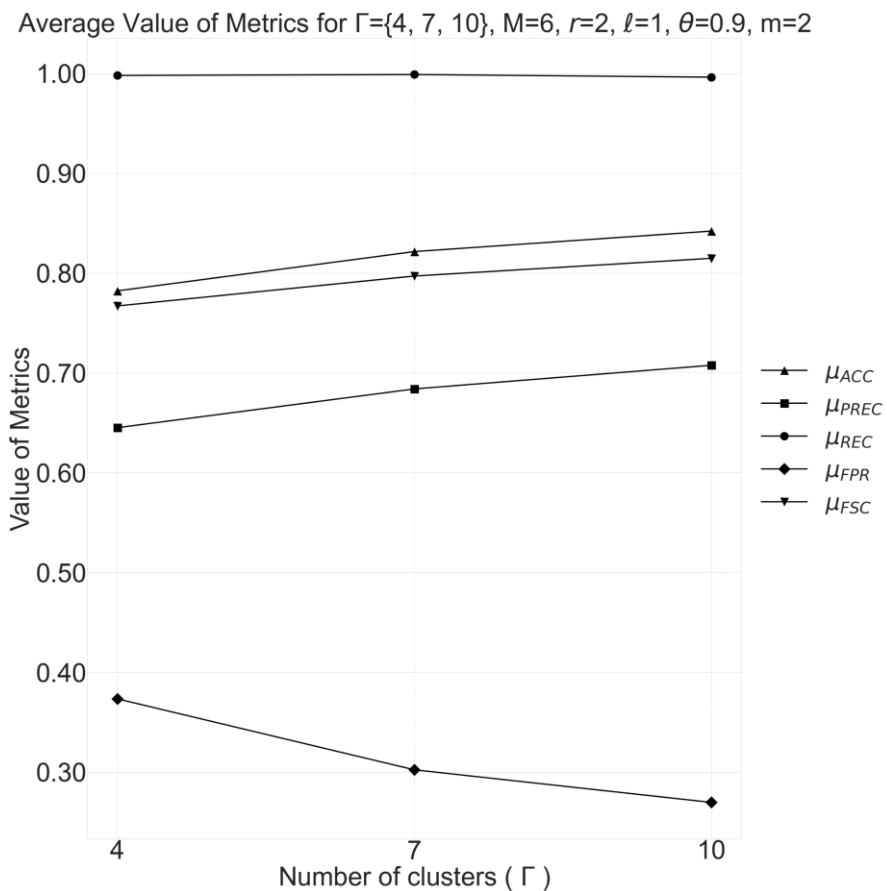
K : the number of clusters which are created in HCBM using K-Means

r : the number of top similar clusters

ℓ : the number of top similar subclusters

θ : the overlap threshold

m : the fuzziness in Fuzzy C-Means



- ❖ Data mapping is a significant data management process in various applications domains.
- ❖ The HMCM can efficiently detect the appropriate data with low error and in short time.
- ❖ Future Research Direction 1: implementation of a more complex methodology for the improvement both of error and time metrics.
- ❖ Future Research Direction 2: adoption of a deep learning model that will be able to adapt our model to changes in user requirements expressed through queries.



THANK YOU

Panagiotis Fountas

Email: pfountas@uth.gr

<http://www.iprism.eu>

