شبكة المعلومات الجامعية

# بسم اللّه الرحمن الرحيم



# شبكة المعلومات الجامعية

# التوثيق الالكتروني والميكروفيلم

شبكة المعلومات الجامعية

# جامعة عين شمس

## التوثيق الإلكتروني والميكروفيلم

# قسم

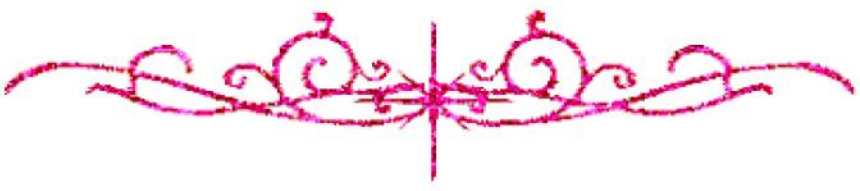
## نقسم باللّه العظيم أن المادة التي تم توثيقها وتسجيلها علي هذه الأقراص المدمجة قد أعدت دون أية تغيرات

# يجب أن

## تحفظ هذه الأقراص المدمجة بعيدا عن الغبار

شبكة المعلومات الجامعية


# بعض الوثائق الأصلية تالفة



# بالرسالة صفحات
# لم ترد بالأصل

Ain Shams University
Faculty of Computer and Information Sciences
Information Systems Department


# An Effective Approach for Processing Data Streams Queries

A thesis submitted in partial fulfillment of the requirements for
the degree of Ph.D in Computer and Information Sciences
To
Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

**By**

**Fatma Mohamed Mahmoud Najib**
Assistant Lecturer, Department of Information Systems
Faculty of Computer and Information Sciences, Ain Shams University

**Under Supervision of**

**Prof. Dr. Tarek Fouad Gharib**
Professor, Head of Information Systems Department,
Faculty of Computer and Information Sciences, Ain Shams University

**Prof. Dr. Nagwa Badr**
Professor, Dean of Faculty of Computer and Information Sciences,
Faculty of Computer and Information Sciences, Ain Shams University

**Assoc. Prof. Rasha Ismail**
Associate Professor, Vice Dean for Graduate Studies and Research,
Faculty of Computer and Information Sciences, Ain Shams University

October – 2020

# Acknowledgement

First of all thanks to Allah for giving me the opportunity and the capability to finish this thesis. Great thanks to everyone who assists and encourages me to complete this work.

I would like to deeply thank Prof. Dr. Tarek Gharib for his leadership, support and encouragement all time.

Also, I would like to thank Prof. Dr. Nagwa Badr for her supervision, instructions and help throughout this thesis.

I would also like to thank Assoc. Prof. Rasha Ismail for her guidance, assistance and support through the work.

Finally, all thanks to my father for being the reason to any success in my life. I would also like to thank every member of my family for their encouragement.

# ABSTRACT

*Most of recent applications such as sensor networks generate continuous and time varying data which are called data streams. Additional constraints are faced for efficient query processing of such data streams that have uncertain nature and require fast and timely processing. Traditional query processing techniques of static data process the whole data without partitioning them, which is not applicable to data streams. Applying data clustering is demanded as a preprocessing step of data streams. Also, data streams are often suffer from incompleteness and high dimensionality. So, in this thesis, we introduce a framework for efficiently answering incomplete high dimensional data streams queries.*

*The proposed framework handles the incompleteness issue by estimating missed values based on the corresponding nearest-neighbors' intervals. The continuous clustering mechanism is adopted and extended to accurately handle the incomplete data streams. The performance of incomplete data clustering was improved by 20% compared to the alternative approaches using two different data sets.*

*The proposed framework provides an improved subspace clustering to deal with high dimensional data streams. The experimental results using two datasets proved the efficiency of the proposed framework on average by 7.9% over the comparing algorithms for clustering such incomplete high dimensional data streams. The query processing performance of the incomplete high dimensional data streams was improved by 62% using two different data sets over the compared algorithms due the proposed clustering improvements of such data.*

# Table of Contents

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| DBMS | Database management systems |
| DMSC | The discriminative multi-view subspace clustering algorithm |
| DS | The dense segments algorithm |
| DSMS | Data streams management systems |
| FCM | The fuzzy c-means algorithm |
| FID | The proposed fuzzy c-means clustering for incomplete data streams framework |
| FNM-T | The fuzzy non-metric model for data with tolerance |
| GBDC-P2P | The gossip-based distributed clustering for P2P networks algorithm |
| GFCM | The Grey based Fuzzy c-Means algorithm |
| IC | The improved clustering algorithm |
| IHDQ | The proposed incomplete high dimensional data streams query processing framework |
| IKFCM | The interval kernel fuzzy c-means clustering for incomplete data algorithm |
| KNN-FS | The KNN-based feature selection algorithm |
| LADEL | The label-aware distributed ensemble learning algorithm |
| LP | The Lattice Partition algorithm |
| MIDC | The multiple imputation of discrete combinations algorithm |
| MIGEC | The multiple imputation using gray-system-theory and entropy based on clustering algorithm |
| MMIC | The model based missing value imputation using correlation method |
| MRBIG | The map reduced enhanced bitmap index guided algorithm |
| MSC_IAS | The multi-view subspace clustering with intactness-aware similarity algorithm |
| OBAL | The online batch-based active learning algorithm |
| PS | The pointed subspace algorithm |
| SC | The smart scheme method |
| SIHD | The proposed subspace clustering for incomplete high dimensional data streams framework |
| SIM | The shape interaction matrix approach |
| SNMF | The symmetric non-negative matrix factorization approach |
| SSR | Sparse sample self-representation |
| VQ_fk_nps | The Gaussian kernel function and the nearest prototype strategy |

# LIST OF PUBLICATIONS

| |
|---|
| [1] Najib, F.M., Ismail, R.M., Badr, N.L. and Gharib, T., 2020. Clustering based approach for incomplete data streams processing. Journal of Intelligent & Fuzzy Systems, 38 (3), 3213-3227. |
| [2] Najib, F.M., Ismail, R.M., Badr, N.L. and Gharib, T., 2020. Incomplete high dimensional data streams clustering. Journal of Intelligent & Fuzzy Systems, 39 (3), 4227-4243. |
| [3] Najib, F.M., Ismail, R.M., Badr, N.L. and Gharib, T., 2020. An efficient approach for query processing of incomplete high dimensional data streams. In: The 7th International Conference on Advanced Machine Learning and Technologies and Applications (AMLTA2021). (Accepted) |