

**ADAPTIVE FIREFLY ALGORITHM FOR HIERARCHICAL
TEXT CLUSTERING**



ATHRAA JASIM MOHAMMED

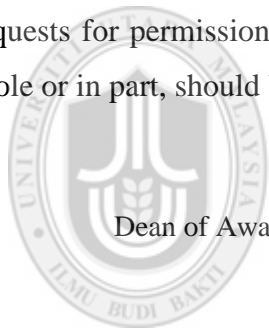
UUM
Universiti Utara Malaysia

**DOCTOR OF PHILOSOPHY
UNIVERSITI UTARA MALAYSIA
2016**

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:



Dean of Awang Had Salleh Graduate School of Arts and Sciences

UUM College of Arts and Sciences

Universiti Utara Malaysia

06010 UUM Sintok

Abstrak

Penggugusan teks digunakan oleh enjin carian untuk meningkatkan recall dan precision dalam bidang capaian maklumat. Memandangkan enjin carian beroperasi menggunakan kandungan Internet yang selalu berubah, maka satu algoritma penggugusan yang menawarkan pengumpulan item secara automatik tanpa maklumat awal berkenaan koleksi berkenaan adalah diperlukan. Kaedah penggugusan sedia ada menghadapi masalah untuk menentukan bilangan gugusan yang optimal dan gugusan yang padat. Dalam penyelidikan ini, satu algoritma penggugusan teks hierarki yang adaptif telah dicadangkan berdasarkan algoritma Firefly. Algoritma Firefly Adaptive (AFA) yang dicadangkan mempunyai tiga komponen: penggugusan dokumen, pembaikan gugusan dan penggabungan gugusan. Komponen pertama memperkenalkan algoritma Weight-based Firefly (WFA) yang berupaya untuk mengenal pasti pusat awalan dan gugusannya secara automatik bagi sesuatu koleksi teks. Bagi memperbaiki gugusan yang telah diperolehi, algoritma kedua iaitu Weight-based Firefly dengan Relocate (WFA_R) telah dicadangkan. Kaedah ini membolehkan penempatan semula dokumen yang telah ditempatkan ke dalam gugusan yang baharu terhasil. Komponen ketiga, Weight-based Firefly Algorithm dengan Relocate dan Merging (WFA_{RM}), bertujuan mengurangkan bilangan gugusan yang terhasil dengan menggabungkan gugusan bukan asli ke dalam gugusan asli. Eksperimen telah dilaksanakan untuk membandingkan algoritma yang dicadangkan dengan tujuh kaedah sedia ada. Peratusan kejayaan memperolehi bilangan gugusan yang optimal oleh AFA ialah 100% dengan mendapat purity dan f-measure 83% lebih tinggi daripada kaedah penanda aras. Bagi ukuran entropy, AFA menghasilkan nilai terendah (0.78) apabila dibandingkan dengan kaedah sedia ada. Keputusan ini memberi indikasi bahawa Algoritma Firefly Adaptif boleh menghasilkan gugusan yang padat. Penyelidikan ini menyumbang kepada domain perlombongan teks memandangkan penggugusan teks hierarki membantu pengindeksan dokumen dan proses pencapaian maklumat.

Kata kunci: Perlombongan teks, Penggugusan teks hierarki, Swarm Intelligence, Firefly Algorithm

Abstract

Text clustering is essentially used by search engines to increase the recall and precision in information retrieval. As search engine operates on Internet content that is constantly being updated, there is a need for a clustering algorithm that offers automatic grouping of items without prior knowledge on the collection. Existing clustering methods have problems in determining optimal number of clusters and producing compact clusters. In this research, an adaptive hierarchical text clustering algorithm is proposed based on Firefly Algorithm. The proposed Adaptive Firefly Algorithm (AFA) consists of three components: document clustering, cluster refining, and cluster merging. The first component introduces Weight-based Firefly Algorithm (WFA) that automatically identifies initial centers and their clusters for any given text collection. In order to refine the obtained clusters, a second algorithm, termed as Weight-based Firefly Algorithm with Relocate (WFA_R), is proposed. Such an approach allows the relocation of a pre-assigned document into a newly created cluster. The third component, Weight-based Firefly Algorithm with Relocate and Merging (WFA_{RM}), aims to reduce the number of produced clusters by merging non-pure clusters into the pure ones. Experiments were conducted to compare the proposed algorithms against seven existing methods. The percentage of success in obtaining optimal number of clusters by AFA is 100% with purity and f-measure of 83% higher than the benchmarked methods. As for entropy measure, the AFA produced the lowest value (0.78) when compared to existing methods. The result indicates that Adaptive Firefly Algorithm can produce compact clusters. This research contributes to the text mining domain as hierarchical text clustering facilitates the indexing of documents and information retrieval processes.

Keywords: Text mining, Hierarchical text clustering, Swarm Intelligence, Firefly Algorithm

Acknowledgement

Firstly, I would like to express my gratitude to Allah (S.W.T.) who helps me to complete my thesis.

Highly appreciate and gratefully acknowledges to my supervisors, Dr. Yuhanis Yusof and Dr. Husniza Husni who they support me, continues encourage me and guides me during my study.

I would like to thank my family for being here with me and supporting me during my study.



Table of Contents

Permission to Use	i
Abstrak.....	ii
Abstract.....	iii
Acknowledgement	iv
Table of Contents.....	v
List of Tables	ix
List of Figures.....	xiii
List of Appendices	xvii
List of Abbreviations	xviii
CHAPTER ONE INTRODUCTION	1
1.1 Research Background.....	4
1.1.1 Clustering	5
1.1.2 Text Clustering.....	7
1.2 Problem Statement	9
1.3 Research Questions	10
1.4 Research Objectives	10
1.5 Research Significance.....	11
1.6 Scope and Limitations of the Research.....	12
1.7 Organization of the Research	12
CHAPTER TWO LITERATURE REVIEW	15
2.1 Introduction.....	15
2.2 Clustering Methods	16
2.2.1 Static Approach.....	16
2.2.1.1 Traditional Methods.....	16
2.2.1.1.1 Partitional Text Clustering	17
2.2.1.1.2 Density-based Text Clustering	22
2.2.1.1.3 Grid-based Text Clustering	25
2.2.1.1.4 Model-based Text Clustering	27
2.2.1.1.5 Hierarchical Text Clustering	29
Agglomerative Clustering	29

Divisive Clustering	35
2.2.1.2 Optimization Methods	42
2.2.1.2.1 Particle Swarm Optimization	46
2.2.1.2.2 Ant Colony Optimization	50
2.2.1.2.3 Firefly Algorithm.....	53
2.2.1.2.4 Hybrid of Clustering Techniques and other Search Optimization	62
2.2.2 Dynamic Approach	66
2.2.2.1 Estimation Approach	66
2.2.2.2 Population-based Approach	68
2.3 Research Gap	74
2.4 Summary	76
CHAPTER THREE RESEARCH METHODOLOGY	78
3.1 Research Design.....	79
3.1.1 Data Acquisition Phase	80
3.1.1.1 Data Collection	81
3.1.1.2 Data Pre-processing	82
Step 1: Data Cleaning	83
Step 2: Data Representation	86
3.1.2 Clustering Phase.....	88
3.1.3 Cluster Refining Phase.....	93
3.1.4 Cluster Merging Phase	95
3.2 Implementation of Algorithms	98
3.3 Evaluation	99
3.3.1 Performance Metrics	100
3.3.1.1 Internal and Relative Quality Metrics	100
3.3.1.2 External Quality Metrics	101
3.3.2 Statistical Analysis	103
3.4 Summary	104
CHAPTER FOUR DOCUMENT CLUSTERING.....	106
4.1 Weight-based Firefly Algorithm (WFA)	106

4.1.1 Initialization of Parameters	106
4.1.2 Data Clustering	108
4.2 Evaluation	119
4.3 Summary	127
CHAPTER FIVE CLUSTER REFINING.....	128
5.1 Introduction	128
5.2 Document Re-locating	128
5.3 Evaluation	132
5.3.1 Comparison between WFA_R and WFA	132
5.3.2 Comparison between WFA_R and Other Methods	140
5.4 Summary	147
CHAPTER SIX CLUSTER MERGING.....	149
6.1 Introduction	149
6.2 Cluster Merging Algorithm.....	150
6.2.1 Merge Clusters	150
6.2.2 Refine Merged Clusters	152
6.3 Evaluation	159
6.3.1 Comparison between WFA_{RM} and WFA_R	160
6.3.1.1 Number of Clusters between WFA_{RM} and WFA_R	160
6.3.1.2 Performance Metrics between WFA_{RM} and WFA_R	161
6.3.1.3 Paired Samples T-test between WFA_{RM} and WFA_R	168
6.3.2 Comparison between WFA_{RM} and Static Methods.....	169
6.3.2.1 Number of Clusters between WFA_{RM} and Static Methods.....	169
6.3.2.2 Performance Metrics between WFA_{RM} and Static Methods.....	170
6.3.2.3 Independent Samples T-test between WFA_{RM} and Static Methods.....	177
6.3.3 Comparison between WFA_{RM} and Dynamic Methods	179
6.3.3.1 Number of Clusters between WFA_{RM} and Dynamic Methods	179
6.3.3.2 Performance Metrics between WFA_{RM} and Dynamic Methods	180
6.3.3.3 Independent Samples T-test between WFA_{RM} and Dynamic Methods	185
6.4 Summary	186

CHAPTER SEVEN EVALUATION OF ADAPTIVE FA ON VARIOUS DATASETS.....	188
7.1 Introduction.....	188
7.2 Comparison WFA _{RM} with Static Methods.....	188
7.2.1 Evaluation Number of Clusters between WFA _{RM} and Static Methods.....	189
7.2.2 Evaluation of Performance Metrics between WFA _{RM} and Static Methods.....	190
7.2.3 Evaluation Independent Samples T-test between WFA _{RM} and Static Methods.....	204
7.3 Comparison WFA _{RM} with Dynamic Methods.....	213
7.3.1 Evaluation Number of Clusters between WFA _{RM} and Dynamic Methods.....	214
7.3.2 Evaluation Performance Metrics between WFA _{RM} and Dynamic Methods.....	215
7.3.3 Evaluation Independent Samples T-test between WFA _{RM} and Dynamic Methods.....	227
7.4 Summary.....	234
CHAPTER EIGHT CONCLUSION AND FUTURE WORK.....	236
8.1 Research Contribution.....	236
8.2 Future Work.....	237

List of Tables

Table 2.1 Summary of existing researches in partitional text clustering.	21
Table 2.2 Summary of existing researches in partitional numerical clustering.	22
Table 2.3 Summary of existing researches in hierarchical text clustering.	40
Table 2.4 Summary of existing researches in hierarchical numerical clustering.	42
Table 2.5 .Summary of existing researches in Particle Swarm Optimization in text clustering.	49
Table 2.6 Summary of existing researches in Particle Swarm Optimization in numerical clustering.	50
Table 2.7 Summary of existing researches in Ant Colony Optimization in text clustering. ...	52
Table 2.8 Summary of existing researches in Ant Colony Optimization in numerical clustering.	53
Table 2.9 Summary of existing researches in Firefly Algorithm in web intelligent data.	61
Table 2.10 Summary of existing researches in Firefly Algorithm in numerical clustering. ..	61
Table 2.11 Summary of existing researches in the hybridization of clustering techniques and other search optimization in text clustering.	65
Table 2.12 Summary of existing researches in the hybridization of clustering techniques and other search optimization in numerical clustering.	65
Table 3.1 Description of Datasets.	81
Table 4.1 Parameters setting in WFA.	111
Table 4.2 External quality metrics of clustering: WFA vs. PSO vs. K-means vs. FAK-means vs. Bisect K-means.	120
Table 4.3 Internal and relative quality metrics of clustering: WFA vs. PSO vs. K-means vs. FAK-means vs. Bisect K-means.	121
Table 4.4 Average number of clusters of WFA vs. PSO vs. K-means vs. FAK-means vs. Bisect K-means.	126
Table 4.5 Results of quality performance of WFA vs. PSO vs. K-means vs. FAK-means vs. Bisect K-means.	127
Table 5.1 External quality metrics: WFA vs. WFA _R	133
Table 5.2 Internal and relative quality metrics: WFA vs. WFA _R	134
Table 5.3 Average number of clusters: WFA vs. WFA _R	139
Table 5.4 Summary of quality performance: WFA vs. WFA _R	139

Table 5.5 ... External quality metrics: WFA _R vs. PSO vs. K-means vs. FAK-means vs. Bisect K-means.	141
Table 5.6 Internal and Relative quality metrics: WFA _R vs. PSO vs. K-means vs. FAK-means vs. Bisect K-means.....	142
Table 5.7 Average number of clusters: WFA _R vs. PSO vs. K-means vs. FA K-means vs. Bisect K-means.	146
Table 5.8 Summary of quality performance: WFA _R vs. PSO vs. K-means vs. FAK-means vs. Bisect K-means.	147
Table 6.1 Average number of clusters of WFA _R & WFA _{RM}	160
Table 6.2 External quality metrics of clustering and standard deviation: WFA _R vs. WFA _{RM}	162
Table 6.3 Internal and relative quality metrics of clustering and standard deviation: WFA _R vs. WFA _{RM}	163
Table 6.4 Quality performance of WFA _R & WFA _{RM} algorithms.	168
Table 6.5 The P-value between WFA _R & WFA _{RM} algorithms.	169
Table 6.6 Average number of clusters: WFA _{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FA K-means vs. BatK-means.....	170
Table 6.7 External quality metrics of clustering: WFA _{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FA K-means vs. BatK-means.....	172
Table 6.8 Internal and relative quality metrics of clustering and standard deviation: WFA _{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FA K-means vs. BatK-means.....	173
Table 6.9 Summary of external quality performance results: WFA _{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FAK-means vs. BatK-means.	176
Table 6.10 Summary of internal and relative quality performance results: WFA _{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FAK-means vs. BatK-means.	177
Table 6.11 The P-value between WFA _{RM} & static methods.....	177
Table 6.12 Average number of clusters: WFA _{RM} vs. PGSCM vs. DCPG.	179
Table 6.13 External quality metrics of clustering and standard deviation: WFA _{RM} vs. PGSCM vs. DCPG.....	180
Table 6.14 Internal and relative quality metrics of clustering and standard deviation: WFA _{RM} vs. PGSCM vs. DCPG.	182
Table 6.15 Summary of quality performance results: WFA _{RM} vs. PGSCM vs. DCPG.....	184
Table 6.16 The P-value between WFA _{RM} & dynamic methods.....	185
Table 7.1 Average numbers of clusters: WFA _{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FAK-means vs. BatK-means using different datasets.	189

Table 7.2 External quality Purity (average, best, worst, standard deviation): WFA _{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FA K-means vs. BatK-means using different datasets (balanced and un-balanced datasets).....	191
Table 7.3 External quality F-measure (average, best, worst, standard deviation): WFA _{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FA K-means vs. BatK-means using different datasets.....	193
Table 7.4 External quality Entropy (average, best, worst, standard deviation): WFA _{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FAK-means vs. BatK-means using different datasets.....	195
Table 7.5 Internal quality ADDC (average, best, worst, standard deviation): WFA _{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FAK-means vs. BatK-means using different datasets.....	197
Table 7.6 Relative quality DBI (Average, Best, Worst, standard deviation): WFA _{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FA K-means vs. BatK-means using different datasets.....	199
Table 7.7 Relative quality DI (average, best DI, worst DI, standard deviation): WFA _{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FA K-means vs. BatK-means using different datasets.....	201
Table 7.8 Summary of quality performance results: WFA _{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FAK-means vs. BatK-means.....	203
Table 7.9 The P-value between WFA _{RM} & static methods using average purity results (sig 2 tailed) with different datasets.....	205
Table 7.10 The P-value between WFA _{RM} & static methods using average F-measure results (sig 2 tailed) with different datasets.....	207
Table 7.11 The P-value between WFA _{RM} & static methods using average Entropy results (sig 2 tailed) with different datasets.....	208
Table 7.12 The P-value between WFA _{RM} & static methods using average ADDC results (sig 2 tailed) with different datasets.....	210
Table 7.13 The P-value between WFA _{RM} & static methods using average DBI results (sig 2 tailed) with different datasets.....	211
Table 7.14 The P-value between WFA _{RM} & static methods using average DI results (sig 2 tailed) with different datasets.....	212
Table 7.15 Average number of clusters: WFA _{RM} vs. PGSCM vs. DCPG using different datasets.....	214

Table 7.16 External quality Purity (average, best, worst, standard deviation): WFA _{RM} vs. PGSCM vs. DCPG using different datasets.	216
Table 7.17 External quality F-measure (average, best, worst, standard deviation): WFA _{RM} vs. PGSCM vs. DCPG using different datasets.	218
Table 7.18 External quality Entropy (average, best, worst, standard deviation): WFA _{RM} vs. PGSCM vs. DCPG using different datasets.	219
Table 7.19 Internal quality ADDC (average, best, worst, standard deviation): WFA _{RM} vs. PGSCM vs. DCPG using different datasets.	221
Table 7.20 Relative quality DBI (average, best, worst, standard deviation): WFA _{RM} vs. PGSCM vs. DCPG using different datasets.	222
Table 7.21 Relative quality DI (average, best DI, worst DI, standard deviation): WFA _{RM} vs. PGSCM vs. DCPG using different datasets.	224
Table 7.22 Summary of quality performance results: WFA _{RM} vs. PGSCM vs. DCPG.	225
Table 7.23 The P-value between WFA _{RM} & dynamic methods using average purity results (sig 2 tailed) with different datasets.	228
Table 7.24 The P-value between WFA _{RM} & dynamic methods using average F-measure results (sig 2 tailed) with different datasets.	229
Table 7.25 The P-value between WFA _{RM} & dynamic methods using average Entropy results (sig 2 tailed) with different datasets.	230
Table 7.26 The P-value between WFA _{RM} & dynamic methods using average ADDC results (sig 2 tailed) with different datasets.	231
Table 7.27 The P-value between WFA _{RM} & dynamic methods using average DBI results (sig 2 tailed) with different datasets.	232
Table 7.28 The P-value between WFA _{RM} & dynamic methods using average DI results (sig 2 tailed) with different datasets.	233

List of Figures

Figure 1.1. Text analytics techniques and external disciplines	3
Figure 2.1. Proposed taxonomy of clustering methods.....	15
Figure 2.2. Steps of K-means algorithm	17
Figure 2.3. The Single Linkage Hierarchical Clustering (SLHC).....	30
Figure 2.4. The Complete Linkage Clustering Hierarchical (CLHC).....	31
Figure 2.5. The Un-weighted Pair Group Method with Arithmetic Mean (UPGMA) Resource. Manning, Raghavan, and Schütze (2008)	32
Figure 2.6. The process of Bisect K-means	36
Figure 2.7. The taxonomy of optimization algorithms	43
Figure 2.8. The step-by-step process of PSO clustering	47
Figure 2.9. Pseudo code of Firefly Algorithm	55
Figure 2.10. Pseudo code of integrated Firefly with K-means clustering algorithm	58
Resource. Tang, Fong, Yang, and Deb (2012).....	58
Figure 2.11. Pseudo code of integrated Bat with K-means clustering algorithm.....	59
Figure 2.12. Pseudo code of integrating Particle Swarm Optimization with Genetic Algorithm (DCPG).....	70
Figure 2.13. Pseudo code of practical General Stochastic Clustering Method (PGSCM).....	73
Figure 3.1. The experimental research steps	78
Figure 3.2. The components of the proposed Adaptive Firefly algorithm for hierarchical text clustering.....	79
Figure 3.3. The phases of proposed hierarchical text clustering.....	80
Figure 3.4. An example of document from the Reuters dataset.....	83
Figure 3.5. An example of a cleaned document.....	83
Figure 3.6. An example of extracted terms	84
Figure 3.7. An example of words with the length more than two.....	84
Figure 3.8: An example of the removed stop words.	85
Figure 3.9. An example of word frequency	85
Figure 3.10. The term frequency matrix	86
Figure 3.11. TFIDF matrix.....	88
Figure 3.12. Flow of Hierarchical Text clustering using Weight-based Firefly Algorithm (WFA).....	89
Figure 3.13. An example of the total weight matrix	90

Figure 3.14. The process of Weight-based Firefly Algorithm (WFA).....	92
Figure 3.15. Process of document re-locating.....	93
Figure 3.16. Comparison between clusters for document re-locating.....	94
Figure 3.17. Process of merging similar clusters in enhanced Un-weighted Pair Group Method with Arithmetic Mean (eUPGMA).....	96
Figure 4.1. One dimension search space.....	107
Figure 4.2. An example of normalized positioning.....	107
Figure 4.3. An example of competition in standard Firefly Algorithm (FA).....	109
Figure 4.4. An example of competition in Weight-based Firefly Algorithm (WFA)	109
Figure 4.5. Weight-based Firefly Algorithm (WFA) for hierarchical text clustering.....	113
Figure 4.6. An example of TFIDF for 20Newsgroups.....	114
Figure 4.7. An example of cosine similarity table for 20Newsgroups dataset.....	115
Figure 4.8. An example of Euclidean distance table for 20Newsgroups dataset.....	115
Figure 4.9. An example of total weight for 20Newsgroups dataset.....	116
Figure 4.10. An example of normalized initial positioning for 20Newsgroups dataset.....	117
Figure 4.11. Graphical representation of initial document positioning for 20Newsgroups dataset.....	117
Figure 4.12. An example of graphical representation of final document positioning for 20Newsgroups dataset.....	118
Figure 4.13. Graphical representation of quality metrics of WFA vs. PSO vs. K-means vs. FAK-means vs. Bisect K-means; a) Purity, b) F-measure, c) Entropy, d) ADDC, e) DBI, and f) DI.....	122
Figure 5.1. The pseudo code of Document Re-locating.....	129
Figure 5.2. The process of WFA_R	129
Figure 5.3. Steps of the WFA_R algorithm	130
Figure 5.4. Graphical representation of quality metrics between WFA & WFA_R ; a) Purity, b) F-measure, c) Entropy, d) ADDC, e) DBI, and f) DI.....	135
Figure 5.5. Graphical representation of quality metrics of WFA_R vs. PSO vs. K-means vs. FAK-means vs. Bisect K-means; a) Purity, b) F-measure, c) Entropy, d) ADDC, e) DBI, and f) DI.....	143
Figure 6.1. Process in WFA_{RM}	149
Figure 6.2. Process of cluster merging Algorithm (eUPGMA)	150
Figure 6.3. Pseudo code for selecting pure clusters	153
Figure 6.4. Pseudo code of identifying centers for pure clusters	153
Figure 6.5. Pseudo code of relocating non-pure clusters	154

Figure 6.6. Cosine similarity matrix between cluster1 and cluster2	155
Figure 6.7. Results of merging clusters for 20Newsgroups dataset	157
Figure 6.8. An example of TFIDF of documents in Cluster1 and center calculation	158
Figure 6.9. An example of TFIDF of document 28 in Cluster 3.....	158
Figure 6.10. An example of the centers of selected pure clusters.....	158
Figure 6.11. Calculation of minimum distance between centers of pure clusters and members of non-pure cluster	159
Figure 6.12. Number of produced clusters by WFA_R and WFA_{RM}	161
Figure 6.14. Graphical representation of quality metrics: WFA_{RM} vs. PSO vs. K- means vs. Bisect K-means vs. FA K-means vs. BatK-means (a) Purity, (b) F-measure, (c) Entropy, (d) ADDC, (e) DBI, and (f) DI.....	174
Figure 6.15. External quality metrics: WFA_{RM} vs. PGSCM vs. DCPG.....	181
Figure 6.16. Internal and relative quality metrics: WFA_{RM} vs. PGSCM vs. DCPG	183
Figure 7.1. Results of the number of generated clusters by WFA_{RM} and the real number of clusters of all static methods	190
Figure 7.2. Average Purity results: WFA_{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FA K-means vs. BatK-means using different datasets.....	192
Figure 7.3. Average F-measure result: WFA_{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FA K-means vs. BatK-means using different datasets.....	194
Figure 7.4. Average Entropy result: WFA_{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FA K-means vs. BatK-means using different datasets.....	196
Figure 7.5. Average ADDC result: WFA_{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FA K-means vs. BatK-means using different datasets.....	198
Figure 7.6. Average DBI result: WFA_{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FA K- means vs. BatK-means using different datasets.....	200
Figure 7.7. Average DI result: WFA_{RM} vs. PSO vs. K-means vs. Bisect K-means vs. FA K- means vs. BatK-means using different datasets.....	202
Figure 7.8. Number of generated clusters: WFA_{RM} vs. the real number of clusters vs. PGSCM vs. DCPG.....	215
Figure 7.9. Average Purity result: WFA_{RM} vs. PGSCM vs. DCPG using different datasets	217
Figure 7.10. Average F-measure result: WFA_{RM} vs. PGSCM vs. DCPG using different datasets.....	218
Figure 7.11. Average Entropy result: WFA_{RM} vs. PGSCM vs. DCPG using different datasets	220

Figure 7.12. Average Entropy result: WFA_{RM} vs. PGSCM vs. DCPG using different datasets 221

Figure 7.13. Average DBI result: WFA_{RM} vs. PGSCM vs. DCPG using different datasets 223

Figure 7.14. Average DI result: WFA_{RM} vs. PGSCM vs. DCPG using different datasets. . 225



List of Appendices

Appendix A Samples of Documents Datasets	253
Appendix B Stop Words List	259



List of Abbreviations

ACK	Ant Colony with Kernal method
ACO	Ant Colony Optimization
ACPSO	Automatic Clustering Particle Swarm Optimization
ALHC	Average Linkage Hierarchical Clustering
AP	Affinity Propagation
BIC	Bayesian Information Criterion
BKM	Bisect K-means
C-bat	Bat algorithm with K-means
C-cuckoo	Cuckoo algorithm with K-means
C-firefly	Firefly algorithm with K-means
CFWS	Clustering based on Frequent Word Sequence
CLHC	Complete Linkage Hierarchical Clustering
CLIQUE	Clustering In QUEst
CMS	Clustering based on Maximal Frequent Sequence
CPSO	Particle Swarm Optimization with K-means
CRC	Corrected Rand Coefficient
C-wolf	Wolf algorithm with K-means
DBI	Davies Bouldin Index
DBSCAN	Density-Based Spatial Clustering of Application with Noise
DCGA	Dynamic Clustering Genetic Algorithm
DCPG	Dynamic Clustering Particle Swarm Optimization with Genetic Algorithm
DCPSO	Dynamic Clustering using Particle Swarm Optimization
DF	Document Frequency
DHC	Dynamic Hierarchical Compact
DHS	Dynamic Hierarchical Star
DI	Dunn Index
ES	Evolution Strategy
FA	Firefly Algorithm
FIHC	Frequent Itemset based Hierarchical Clustering
FTC	Frequent Term based Clustering
GA	Genetic Algorithm

GGCA	General Grid Clustering Approach
GSA	Gravitational Search Algorithm
GSA-KM	Gravitational Search Algorithm with K-means
HBMO	Honey Bee Mating Optimization
HCM	Hierarchical Clustering Method
HS	Harmony Search
IDF	Inverse Document Frequency
KCPSO	K-means with Particle Swarm Optimization
KFA	K-means with Firefly Algorithm
KHM	K-Harmonic Means algorithm
KPSO	K-means with Particle Swarm Optimization
NMI	Normalized mutual information
NN	Neural Networks
OptiGrid	Optimal Grid clustering
PDDP	Principal Direction Divisive Partitioning
PGSCM	Practical General Stochastic Clustering Method
PSO	Particle Swarm Optimization
PSOKHM	Particle Swarm Optimization with K-Harmonic Means
RFA	Reachback Firefly Algorithm
SA	Simulated Annealing
SAP	Seed Affinity Propagation
SLHC	Single Linkage Hierarchical Clustering
SOM	Self Organizing Map
STING	Statistical Information Grid-based method
TC	Term Contribution
TFIDF	Term Frequency–Inverse Document Frequency
TSP	Travelling Salesman Problem
UPGMA	Un-weighted Pair Group Method with Arithmetic Mean
VI	Validity Index
VSM	Vector Space Model
WFA	Weight-based Firefly Algorithm
WFA_R	Weight-based Firefly Algorithm with relocating
WFA_{RM}	Weight-based Firefly Algorithm with relocating with merging algorithm

CHAPTER ONE

INTRODUCTION

Adaptation in computer science is the process of a system. Adaptive system adapts its behavior to users depending on the information that can be collected from users and the environment. An adaptive system is a set of entities that interact between them and change their behavior in response to their environment. The aim of adaptive change is to achieve the goal. Artificial systems, such as robots, can adapt with the environment by sensing the new condition through the use of feedback loops (i.e. the output of the system becomes input). Furthermore, it can adapt a parameter from the environment based on the change of the conditions; for example, a new adaptive parameter (speed) changes based on the color of the agent added in the adaptive flocking algorithm (Folino, Forestiero, & Spezzano, 2009), and the value of pheromone at each location introduced in the picking and dropping probability functions of the adaptive ant colony clustering algorithm, and it also improves the similarity scaling factor by automatic adoption (El-Feghi, Errateeb, Ahmadi, & Sid-Ahmed, 2009). The adaptive system utilizes machine learning to adapt its behavior over time (Glass, 2011). Swarm Intelligence provides a useful paradigm for implementing adaptive systems (Kennedy & Eberhart, 2001).

Swarm Intelligence or Swarm Computing is “the emergent collective intelligence of groups of simple agents” (Bonabeau, Dorigo, & Theraulaz, 1999). It is useful to solve some problems that cannot be processed using traditional methods. It is used to find optimal solutions in hard problems, such as Travelling Salesman Problem (TSP)

The contents of
the thesis is for
internal user
only

REFERENCES

- 20NewsgroupsDataSet. (2006). <http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-4/text-learning/www/datasets.html>.
- Abshouri, A. A., & Bakhtiary, A. (2012). A new clustering method based on Firefly and KHM. *Journal of Communication and Computer*, 9, 387–391. Retrieved from <http://www.davidpublishing.com/davidpublishing/Upfile/6/4/2012/2012060483417489.pdf>
- Adaniya, M. H. A. C., Abr̃ao, T., & Proenc,a Jr., M. L. (2013). Anomaly Detection Using Metaheuristic Firefly Harmonic Clustering. *Journal of Networks*, 8(1), 82–91. Retrieved from doi:10.4304/jnw.8.1.82-91
- Aggarwal, C. C., & Zhai, C. X. (2012). A survey of text clustering algorithms. In *In Mining Text Data, Springer US* (pp. 77–128). Retrieved from doi:10.1007/978-1-4614-3223-4_4
- Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). automatic subspace clustering of high dimensional data. *SIGMOD '98 Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, 94–105. Retrieved from doi: 10.1145/276304.276314
- Aliguliyev, R. M. (2009a). Clustering of document collection-A weighted approach. *Elsevier, Expert Systems with Applications*, 36(4), 7904–7916. Retrieved from doi: 10.1016/j.eswa.2008.11.017
- Aliguliyev, R. M. (2009b). Performance evaluation of density-based clustering methods. *Elsevier, Information Sciences*, 179(20), 3583–3602. Retrieved from doi: 10.1016/j.ins.2009.06.012
- Aljanabi, A. I. (2010). *Interacted multiple ant colonies for search stagnation problem. College of Arts and Sciences. Universiti Utara Malaysia.*
- Alsmadi, M. K. (2014). A hybrid firefly algorithm with fuzzy-c mean algorithm for MRI brain segmentation. *American Journal of Applied Sciences*, 11(9), 1676–1691.
- Amigo, E., Gonzalo, J., Artiles, J., & Verdejo, F. (2009). *A comparison of extrinsic clustering evaluation metrics based on formal constraints. Springer, Information Retrieval* (Vol. 12, pp. 461–486). Retrieved from doi: 10.1007/s10791-008-9066-8
- Anitha Elavarasi, S., Akilandeswari, J., & Sathiyabhama, B. (2011). A survey on partition clustering algorithms. *International Journal of Enterprise Computing*

and *Business Systems*, 1(1). Retrieved from Retrieved from at <http://www.ijecbs.com>

Apostolopoulos, T., & Vlachos, A. (2011). Application of the Firefly Algorithm for Solving the Economic Emissions Load Dispatch Problem. *International Journal of Combinatorics*, Volume 201, 23 pages. Retrieved from doi:10.1155/2011/523806

Bache, K., & Lichman, M. (2013). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine: CA: University of California, School of Information and Computer Science.

Banati, H., & Bajaj, M. (2013). Performance analysis of Firefly algorithm for data clustering. *Int. J. Swarm Intelligence*, 1(1), 19–35.

Beasley, D., Bull, D. R., & Martin, R. R. (1993). An Overview of Genetic Algorithms : Part 1, Fundamentals. *University Computing*, 15(2), 58–69.

Bojic, I., Podobnik, V., Ljubi, I., Jezic, G., & Kusek, M. (2012). A self-optimizing mobile network: Auto-tuning the network with firefly-synchronized agents. *Elsevier, Information Sciences*, 182(1), 77–92.

Boley, D. (1998). Principal Direction Divisive Partitioning. *ACM, Data Mining and Knowledge Discovery*, 2(4), 325–344. Retrieved from doi: 10.1023/A:1009740529316

Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm Intelligence: From Natural to Artificial Systems*. New York, NY: Oxford University Press, Santa Fe Institute Studies in the Sciences of Complexity.

Bordogna, G., & Pasi, G. (2012). A quality driven Hierarchical Data Divisive Soft Clustering for information retrieval. *Elsevier, Knowledge-Based Systems*, 26, 9–19. Retrieved from doi:10.1016/j.knosys.2011.06.012

Boussaïd, I., Lepagnot, J., & Siarry, P. (2013). A survey on optimization metaheuristics. *Elsevier, Information Sciences*, 237, 82–117.

Cao, D., & Yang, B. (2010). An improved k-medoids clustering algorithm. In *The 2nd International Conference on Computer and Automation Engineering (ICCAE)* (Vol. 3, pp. 132–135). Singapore: IEEE. Retrieved from doi: 10.1109/ICCAE.2010.5452085

Chehreghani, M. H., Abolhassani, H., & Chehreghani, M. H. (2008). Improving density-based methods for hierarchical clustering of web pages. *Elsevier, Data & Knowledge Engineering*, 67(1), 30–50. Retrieved from doi: 10.1016/j.datak.2008.06.006

- Chen, T. S., Tsai, T. H., Chen, Y. T., Lin, C. C., Chen, R. C., Li, S. Y., & Chen, H. Y. (2005). A combined K-means and hierarchical clustering method for improving the clustering efficiency of microarray. In *Proceedings of intelligent signal processing and communication systems, IEEE* (pp. 405–408). IEEE. Retrieved from doi: 10.1109/ISPACS.2005.1595432
- Cui, X., Gao, J., & Potok, T. E. (2006). A flocking based algorithm for document clustering analysis. *Journal of Systems Architecture*, 52(8-9), 505–515.
- Cui, X., Potok, T. E., & Palathingal, P. (2005). Document Clustering using Particle Swarm Optimization. In *Proceedings 2005 IEEE Swarm Intelligence Symposium, SIS 2005*. (pp. 185–191). IEEEExplore. Retrieved from doi:10.1109/SIS.2005.1501621
- Das, S., Abraham, A., & Konar, A. (2008). Automatic Clustering Using an Improved Differential Evolution Algorithm. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 38(1), 218–237. doi:10.1109/TSMCA.2007.909595
- Das, S., Abraham, A., & Konar, A. (2009). *Metaheuristic Clustering*. Verlag Berlin Heidelberg: Springer. Retrieved from doi: 10.1007/978-3-540-93964-1
- Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Vol. PAMI-1, pp. 224–227). Retrieved from doi:10.1109/TPAMI.1979.4766909
- Demir, M., & Karci, A. (2015). Data Clustering on Breast Cancer Data Using Firefly Algorithm with Golden Ratio Method. *Advances in Electrical and Computer Engineering*, 15(2), 75–84. doi:10.4316/AECE.2015.02010
- Deneubourg, J. L., Goss, S., Franks, N., Sendova-Franks, A., Detrain, C., & Chrétien, L. (1991). The dynamics of collective sorting: robot-like ants and ant-like robots. In *Proceedings of the first international conference on simulation of adaptive behavior on From animals to animats* (pp. 356–363). MIT Press Cambridge, MA, USA.
- Ding, Y., & Fu, X. (2012). The Research of Text Mining Based on Self-Organizing Maps. *Procedia Engineering*, 29(0), 537–541. doi:http://dx.doi.org/10.1016/j.proeng.2011.12.757
- Doding, G. (2002). *Computer Science in a Theory of Science Discourse*. Department of Computer Science. Malardalen University, Sweden.
- Dorigo, M. (1992). *Optimization, Learning and Natural Algorithms*. Politecnico di Milano, Italie.

- Dorigo, M., & Gambardella, L. M. (1997). Ant colonies for the traveling salesman problem. *Elsevier, Biosystems*, 43(2), 73–81. Retrieved from doi:10.1016/S0303-2647(97)01708-5
- Dos Santos Coelho, L., de Andrade Bernert, D. L., & Mariani, V. C. (2011). A chaotic firefly algorithm applied to reliability-redundancy optimization. In *2011 IEEE Congress on Evolutionary Computation (CEC)* (pp. 517–521). New Orleans, LA. Retrieved from doi:10.1109/CEC.2011.5949662
- Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4, 95–104. Retrieved from doi:10.1080/01969727408546059
- El-Abd, M., & Kamel, M. (2005). A taxonomy of cooperative search algorithms. *Hybrid Metaheuristics*, 3636, 32–41. Retrieved from doi:10.1007/11546245_4
- El-Feghi, I., Errateeb, M., Ahmadi, M., & Sid-Ahmed, M. a. (2009). An adaptive ant-based clustering algorithm with improved environment perception. In *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics* (pp. 1431–1438). doi:10.1109/ICSMC.2009.5346291
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. AAAI Press. (pp. 226–231).
- Falcon, R., Almeida, M., & Nayak, A. (2011). Fault Identification with Binary Adaptive Fireflies in Parallel and Distributed Systems. In *2011 IEEE Congress on Evolutionary Computation (CEC)*, (pp. 1359–1366). New Orleans, LA: IEEE Explore. Retrieved from doi:10.1109/CEC.2011.5949774
- Feng, L., Qiu, M. H., Wang, Y. X., Xiang, Q. L., Yang, Y. F., & Liu, K. (2010). A fast divisive clustering algorithm using an improved discrete particle swarm optimizer. *Elsevier, Pattern Recognition Letters*, 31(11), 1216–1225. Retrieved from doi: 10.1016/j.patrec.2010.04.001
- Fister, I., Jr, I. F., Yang, X. S., & Brest, J. (2013). A comprehensive review of Firefly Algorithms. *Elsevier, Swarm and Evolutionary Computation*, 13, 34–46.
- Folino, G., Forestiero, A., & Spezzano, G. (2009). An adaptive flocking algorithm for performing approximate clustering. *Information Sciences*, 179(18), 3059–3078.
- Fong, S., Deb, S., Yang, X. S., & Zhuang, Y. (2014). Towards Enhancement of Performance of K-Means Clustering Using Nature-Inspired Optimization Algorithms. *The Scientific World Journal*, 2014(564829), 16 pages.

- Forsati, R., Mahdavi, M., Shamsfard, M., & Meybodi, M. R. (2013). Efficient stochastic algorithms for document clustering. *Elsevier, Information Sciences*, 220, 269–291. Retrieved from doi: 10.1016/j.ins.2012.07.025
- Gil-Garcia, R., & Pons-Porrata, A. (2010). Dynamic hierarchical algorithms for document clustering. *Elsevier, Pattern Recognition Letters*, 31(6), 469–477. Retrieved from doi: 10.1016/j.patrec.2009.11.011
- Glass, A. (2011). *Explanation of Adaptive Systems*. Stanford University.
- Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research*, 13(No.5), 533–549.
- Gu, J., Zhou, J., & Chen, X. (2009). An Enhancement of K-means Clustering Algorithm. In *IEEE, International Conference on Business Intelligence and Financial Engineering* (pp. 237–240). Beijing: IEEE. Retrieved from doi: 10.1109/BIFE.2009.204
- Guan, R., Shi, X., Marchese, M., Yang, C., & Liang, Y. (2011). Text Clustering with Seeds Affinity Propagation. *IEEE Transactions on Knowledge and Data Engineering*, 23(4), 627–637. Retrieved from doi: 10.1109/TKDE.2010.144
- Gupta, P., & Sharma, A. K. (2010). A framework for hierarchical clustering based indexing in search engines. In *Proceedings of 1st International Conference on Parallel, Distributed and Grid Computing (PDGC - 2010)* (pp. 372–377). Solan: IEEE. Retrieved from doi: 10.1109/PDGC.2010.5679966
- Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques (2nd ed.)*. San Francisco: Morgan Kaufman.
- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques, 3rd edition. The Morgan Kaufmann Series in Data Management Systems* (p. 744 pages). Morgan Kaufmann.
- Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *JStor, Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(No.1). Retrieved from <http://www.jstor.org/stable/2346830>
- Hassanzadeh, T., Faez, K., & Seyfi, G. (2012). A Speech Recognition System Based on Structure Equivalent Fuzzy Neural Network Trained by Firefly Algorithm. In *International Conference on Biomedical Engineering (ICoBE)* (pp. 63–67). Penang: IEEE Explore. Retrieved from doi:10.1109/ICoBE.2012.6178956
- Hassanzadeh, T., & Meybodi, M. R. (2012). A new hybrid approach for data clustering using Firefly algorithm and k-means. In *The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012), IEEE* (pp. 7–11). Retrieved from doi: 10.1109/AISP.2012.6313708

- Hassanzadeh, T., Vojodi, H., & Moghadam, A. M. E. (2011). An Image Segmentation Approach Based on Maximum Variance Intra-Cluster Method and Firefly Algorithm. In *Seventh International Conference on Natural Computation (ICNC)* (Vol. 3, pp. 1817–1821). Shanghai: IEEE Explore. Retrieved from doi:10.1109/ICNC.2011.6022379
- Hatamlou, A., Abdullah, S., & Nezamabadi-pour, H. (2012). A combined approach for clustering based on K-means and gravitational search algorithms. *Elsevier, Swarm and Evolutionary Computation*, 6, 47–52. Retrieved from doi: 10.1016/j.swevo.2012.02.003
- He, Y., Hui, S. C., & Sim, Y. (2006). Anovel ant-based clustering approach document clustering. *Information Retrieval Technology*, 4182, 537–544.
- Hinneburg, A., & Keim, D. (1999). Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering. In *Proceedings of the 25th International Conference on Very Large Data Bases* (pp. 506–517). Morgan Kaufmann Publishers Inc.
- Holland, J. (1992). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence* (p. 211). Cambridge, MA, USA.
- Hornig, M. H., & Jiang, T. W. (2010). Multilevel Image Thresholding Selection based on the Firefly Algorithm. In *7th International Conference on Ubiquitous Intelligence & Computing and 7th International Conference on Autonomic & Trusted Computing (UIC/ATC)*, (pp. 58–63). Xian, Shaanxi: IEEE Explore. Retrieved from doi:10.1109/UIC-ATC.2010.47
- Hu, G., Zhou, S., Guan, J., & Hu, X. (2008). Towards effective document clustering: A constrained K-means based approach. *Elsevier, Information Processing & Management*, 44(4), 1397–1409. Retrieved from doi: 10.1016/j.ipm.2008.03.001
- Ilango, M., & Mohan, V. (2010). A Survey of Grid Based Clustering Algorithms. *International Journal of Engineering Science and Technology*, 2(8), 3441–3446.
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Elsevier, Pattern Recognition Letters*, 31(8), 651–666. Retrieved from doi: 10.1016/j.patrec.2009.09.011
- Jensi, R., & Jiji, D. G. W. (2013). A Survey on optimization approaches to text document clustering. *International Journal on Computational Sciences & Applications (IJCSA)*, 3(6), 31–44.

- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transaction of the ASME-Journal of Basic Engineering*, 82 (series, 35–45).
- Kao, Y., & Lee, S.-Y. (2009). Combining K-means and particle swarm optimization for dynamic data clustering problems. In *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on* (Vol. 1, pp. 757–761).
- Karypis, G. (2002). *CLUTO a clustering toolkit*, Technical Report 02-017. Dept. of Computer Science, University of Minnesota. Retrieved from Available at <http://glaros.dtc.umn.edu/gkhome/views/cluto>
- Karypis, G., Han, E. H., & Kumar, V. (1999). Chameleon: Hierarchical Clustering Using Dynamic Modeling. *IEEE Computer Society*, 32(8), 68–75. Retrieved from doi: 10.1109/2.781637
- Kashef, R., & Kamel, M. (2010). Cooperative clustering. *Elsevier, Pattern Recognition*, 43(6), 2315–2329. Retrieved from doi: 10.1016/j.patcog.2009.12.018
- Kashef, R., & Kamel, M. S. (2009). Enhanced bisecting k-means clustering using intermediate cooperation. *Elsevier, Pattern Recognition*, 42(11), 2557–2569.
- Kennedy, J., & Eberhart, R. (1995). Particle Swarm Optimization. *Proceedings of IEEE International Conference on Neural Networks IV*. Perth, WA: IEEE. Retrieved from doi:10.1109/ICNN.1995.488968
- Kennedy, J. F., & Eberhart, R. C. (2001). *Swarm intelligence* (p. 512). Morgan Kaufmann.
- Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by Simulated Annealing. *New Series*, 220(No. 4598), 671–680.
- Kohonen, T. (1998). The self-organizing map. *Elsevier, Neurocomputing*, 21(1-3), 1–6. Retrieved from doi: 10.1016/S0925-2312(98)00030-7
- Kohonen, T. (2001). *Self organizing map 3rd ed*. Springer-Verlag Berlin Heidelberg NewYork.
- Kuo, R. J., Syu, Y. J., Chen, Z., & Tien, F. C. (2012). Integration of particle swarm optimization and genetic algorithm for dynamic clustering. *Elsevier, Information Sciences*, 195, 124–140.
- Kuo, R. J., & Zulvia, F. E. (2013). automatic clustering using an improved particle swarm optimization. *Journal of Industrial and Intelligent Information*, 1(1), 46–51.

- Lahane, S. V., Kharat, M. U., & Halgaonkar, P. S. (2012). Divisive approach of Clustering for Educational Data. In *Fifth International Conference on Emerging Trends in Engineering and Technology* (pp. 191–195). Himeji: IEEE. Retrieved from doi:10.1109/ICETET.2012.55
- Lee, C. Y., & Antonsson, E. K. (2000). Dynamic partitional clustering using evolution strategies. In *IEEE* (Vol. 4, pp. 2716–2721).
- Lewis, D. (1999). The reuters-21578 text categorization test collection. Retrieved from Available online at :<http://kdd.ics.uci.edu/database/reuters21578/reuters21578.html>
- Liu, Y. C., Wu, C., & Liu, M. (2011). Research of fast SOM clustering for text information. *Elsevier, Expert Systems with Applications*, 38(8), 9325–9333. Retrieved from doi: 10.1016/j.eswa.2011.01.126
- Liu, Y. C., Wu, X., & Shen, Y. (2011). Automatic clustering using genetic algorithms. *Elsevier, Applied Mathematics and Computation*, 218(4), 1267–1279.
- Lu, Y., Wang, S., Li, S., & Zhou, C. (2009). Text Clustering via Particle Swarm Optimization. In *Swarm Intelligence Symposium, 2009. SIS '09. IEEE* (pp. 45–51). Nashville, TN: IEEEXplore. Retrieved from doi:10.1109/SIS.2009.4937843
- Luo, C., Li, Y., & Chung, S. M. (2009). Text document clustering based on neighbors. *Elsevier, Data & Knowledge Engineering*, 68(11), 1271–1288. Retrieved from doi: 10.1016/j.datak.2009.06.007
- MacQueen, J. B. (1967). Kmeans Some Methods for classification and Analysis of Multivariate Observations. *5th Berkeley Symposium on Mathematical Statistics and Probability 1967*, 1(233), 281–297. doi:citeulike-article-id:6083430
- Mahmuddin, M. (2008). *Optimisation using Bees algorithm on unlabelled data problems*. Manufacturing engineering centre. Cardiff university, Cardiff, UK.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval, 1 ed.* New York, USA: Cambridge University Press.
- Martens, D., Backer, M. D., Haesen, R., Vanthienen, J., Snoeck, M., & Baesens, B. (2007). Classification With Ant Colony Optimization. *IEEE Transactions on Evolutionary Computation*, 11(5), 651–665. Retrieved from doi: 10.1109/TEVC.2006.890229
- Meghabghab, G., & Kandel, A. (2008). *Search engines, link analysis, and user's web behaviour* (Vol. 99). Springer Berlin Heidelberg. Retrieved from doi:10.1007/978-3-540-77469-3

- Miner, G., Elder, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, 1st ed.* Elsevier.
- Mishra, B. K., Nayak, N. R., Rath, A., & Swain, S. (2012). Far Efficient K-Means Clustering Algorithm. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics* (pp. 106–110). ACM. Retrieved from doi:10.1145/2345396.2345414
- Muñoz, D. M., Llanos, C. H., Coelho, L. D. S., & Ayala-Rincon, M. (2011). Opposition-based shuffled PSO with passive congregation applied to FM matching synthesis. In *2011 IEEE Congress on Evolutionary Computation (CEC)*, (pp. 2775–2781). New Orleans, LA: IEEE Xplore. Retrieved from doi:10.1109/CEC.2011.5949966
- Murugesan, K., & Zhang, J. (2011a). Hybrid Bisect K-means clustering algorithm. In *International Conference on Business Computing and Global Informatization* (pp. 216–219). Retrieved from doi:10.1109/BCGIn.2011.62
- Murugesan, K., & Zhang, J. (2011b). *Hybrid hierarchical clustering: An experimental analysis* (p. 26). university of Kentucky.
- Nandy, S., Sarkar, P. P., & Das, A. (2012). Analysis of a Nature Inspired Firefly Algorithm based Back-propagation Neural Network Training. *International Journal of Computer Applications*, 43(22), 8–16. Retrieved from doi:10.5120/6401-8339
- Pelleg, M., & Moore, A. (2000). X-means: Extending K-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning* (pp. 727–734). Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.
- Picarougne, F., Azzag, H., Venturini, G., & Guinot, C. (2007). A New Approach of Data Clustering Using a Flock of Agents. *Evolutionary Computation, Cambridge: MIT Press (2007)*, 15(3), 345–367.
- Poomagal, S., & Hamsapriya, T. (2011). Optimized k-means clustering with intelligent initial centroid selection for web search using URL and tag contents. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics* (pp. 1–8). Sogndal, Norway: ACM. Retrieved from doi:10.1145/1988688.1988764
- Pop, C. B., Chifu, V. R., Salomie, I., Baico, R. B., Dinsoreanu, M., & Copil, G. (2011). A Hybrid Firefly-inspired Approach for Optimal Semantic Web Service Composition. *Scientific International Journal for Parallel and Distributed Computing*, 12(3), 363–369. Retrieved from <http://www.scpe.org/index.php/scpe/article/view/730/0>

- Rafsanjani, M. K., Varzaneh, Z. A., & Chukanlo, N. E. (2012). A survey of hierarchical clustering algorithms. *The Journal of Mathematics and Computer Science, TJMCS*, 5, No. 3, 229–240. Retrieved from Available online at: <http://www.TJMCS.com>
- Rana, S., Jasola, S., & Kumar, R. (2010). A hybrid sequential approach for data clustering using K-Means and particle swarm optimization algorithm. *International Journal of Engineering, Science and Technology*, 2, No.6, 167–176. Retrieved from Available online at: <http://www.ajol.info/index.php/ijest/article/view/63708>
- Rashedi, E., Nezamabadi-pour, H., & Saryazdi, S. (2009). GSA: A Gravitational Search Algorithm. *Elsevier, Information Sciences*, 179(13), 2232–2248.
- Rokach, L., & Maimon, O. (2005). *Clustering Methods, Data Mining and Knowledge Discovery Handbook*. Springer (pp. 321–352.).
- Ross, S. M. (2010). *Introductory Statistics*. Elsevier Science. Retrieved from <http://books.google.com.my/books?id=ZKswvkqhygYC>
- Rothlauf, F. (2011). *Design of Modern Heuristics Principles and Application*. Springer-Verlag Berlin Heidelberg. Retrieved from doi:10.1007/978-3-450-72962-4
- Rui, T., Fong, S., Yang, X. S., & Deb, S. (2012). Nature-Inspired Clustering Algorithms for Web Intelligence Data. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (Vol. 3, pp. 147–153). Macau. Retrieved from doi:10.1109/WI-IAT.2012.83
- Sander, J. (2010). Density-Based Clustering. *Encyclopedia of Machine Learning SE - 211*. Springer US DA - 2010/01/01. Retrieved from doi:10.1007/978-0-387-30164-8_211
- Sarkar, M., Yegnanarayana, B., & Khemani, D. (1997). A clustering algorithm using an evolutionary programming-based approach. *Elsevier, Pattern Recognition*, 18(10), 975–986.
- Sayadi, M. K., Hafezalkotob, A., & Naini, S. G. J. (2013). Firefly-inspired algorithm for discrete optimization problems: An application to manufacturing cell formation. *Elsevier, Journal of Manufacturing Systems*, 32(1), 78–84.
- Sayed, A., Hacid, H., & Zighed, D. (2009). Exploring validity indices for clustering textual data. In *Mining Complex Data*, 165, 281–300.
- Senthilnath, J., Omkar, S. N., & Mani, V. (2011). Clustering using firefly algorithm: Performance study. *Elsevier, Swarm and Evolutionary Computation*, 1(3), 164–171. Retrieved from doi: 10.1016/j.swevo.2011.06.003

- Shannon, C. E. (1948). A Mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423, 623–656,. Retrieved from Retrieved from: <http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf>
- Singh, R. V., & Bhatia, M. P. S. (2011). Data clustering with modified K-means algorithm. In *International Conference on Recent Trends in Information Technology (ICRTIT)* (pp. 717–721). Chennai, Tamil Nadu: IEEE. Retrieved from doi:10.1109/ICRTIT.2011.5972376
- Stahlbock, R., Crone, S. F., & Lessmann, S. (2010). *Data Mining Special Issue in Annals of Information Systems* (Vol. 8). Springer US. Retrieved from doi: 10.1007/978-1-4419-1280-0
- Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Pearson Education, Addition Wesley.
- Tan, S. C. (2012). Simplifying and improving swarm based clustering. In *IEEE Congress on Evolutionary Computation (CEC)* (pp. 1–8). Brisbane, QLD: IEEE.
- Tan, S. C., Ting, K. M., & Teng, S. W. (2011a). A general stochastic clustering method for automatic cluster discovery. *Elsevier, Pattern Recognition*, 44(10-11), 2786–2799.
- Tan, S. C., Ting, K. M., & Teng, S. W. (2011b). Simplifying and improving ant-based clustering. In *Procedia computer science* (pp. 46–55).
- Tang, R., Fong, S., Yang, X. S., & Deb, S. (2012). Integrating nature-inspired optimization algorithms to K-means clustering. In *Seventh International Conference on Digital Information Management (ICDIM), 2012* (pp. 116–123). Macau: IEEE. Retrieved from doi:10.1109/ICDIM.2012.6360145
- Toreini, E., & Mehrnejad, M. (2011). Clustering Data with Particle Swarm Optimization Using a New Fitness. In *2011 3rd Conference on Data Mining and Optimization (DMO)* (pp. 266–270). Putrajaya: IEEEExplore. Retrieved from doi:10.1109/DMO.2011.5976539
- TREC. (1999). Text REtrieval Conference (TREC). Retrieved from Available online at :<http://trec.nist.gov/>
- Van der Merwe, D. W., & Engelbrecht, A. P. (2003). Data clustering using particle swarm optimization. In *The 2003 Congress on Evolutionary Computation, 2003. CEC '03.* (Vol. 1, pp. 215–220). Retrieved from doi:10.1109/CEC.2003.1299577
- Vijayalakshmi, M., MCA, M., & Devi, M. R. (2012). A survey of different issue of different clustering algorithms used in large data sets. *International Journal of*

Advance Research in Computer Science and Software Engineering, 2(3), 305–307. Retrieved from Available online at : <http://www.ijarcse.com>

- Wang, H., Yang, X., Zhang, J., Zhang, M., Bai, X., Yin, W., & Dong, J. (2011). BP neural network model based on cluster analysis for wind power prediction. In *2011 IEEE International Conference on Service Operations, Logistics, and Informatics (SOLI)* (pp. 278–280). Beijing: IEEE Xplore. Retrieved from doi:10.1109/SOLI.2011.5986570
- Wang, W., Yang, J., & Muntz, R. (1997). STING : A Statistical Information Grid Approach to Spatial Data Mining. In *VLDB '97 Proceedings of the 23rd International Conference on Very Large Data Bases* (pp. 186–195). Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.
- Wang, X., Shen, J., & Tang, H. (2009). Novel hybrid document clustering algorithm based on Ant Colony and agglomerate. In *Second International Symposium on Knowledge Acquisition and Modeling* (Vol. 3, pp. 65–68). Wuhan: IEEE computer society. Retrieved from doi:10.1109/KAM.2009.182
- Wang, Z., Liu, Z., Chen, D., & Tang, K. (2011). A New Partitioning Based Algorithm For Document Clustering. In *Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)* (Vol. 3, pp. 1741–1745). Shanghai: IEEE. Retrieved from doi: 10.1109/FSKD.2011.6019857
- Wilson, H. G., Boots, B., & Millward, A. A. (2002). A comparison of hierarchical and partitional clustering techniques for multispectral image classification. In *IEEE* (Vol. 3, pp. 1624–1626). IEEE International Geoscience and Remote Sensing Symposium, 2002. IGARSS. Retrieved from doi: 10.1109/IGARSS.2002.1026201
- Xinwu, L. (2010). Research on Text Clustering Algorithm Based on Improved K-means. In *International Conference On Computer Design And Applications (ICCDA 2010)* (Vol. 4, pp. V4–573 – V4–576). Qinhuangdao: IEEE. Retrieved from doi: 10.1109/ICCDA.2010.5540727
- Xu, Y. (2005). Hybrid clustering with application to web mining. In *Proceedings of the International Conference on Active Media Technology (AMT 2005)*. (pp. 574–578). Retrieved from doi: 10.1109/AMT.2005.1505425
- Yang, H. (2010). A Document Clustering Algorithm for Web Search Engine Retrieval System. In *International Conference on e-Education, e-Business, e-Management, and e-Learning, 2010. IC4E '10* (pp. 383–386). Sanya: IEEE. Retrieved from doi:10.1109/IC4E.2010.72
- Yang, X. S. (2009). Firefly Algorithms for Multimodal Optimization. In O. Watanabe & T. Zeugmann (Eds.), *Stochastic Algorithms: Foundations and*

Applications (pp. 169–178). Springer Berlin Heidelberg. doi:10.1007/978-3-642-04944-6_14

- Yang, X. S. (2010a). Firefly Algorithm, Stochastic Test Functions and Design Optimisation. *Int. J. Bio-Inspired Computation*, 2(2), 78–84.
- Yang, X. S. (2010b). *Nature-inspired metaheuristic algorithms 2nd edition*. United Kingdom: Luniver press.
- Yang, X. S., & He, X. (2013). Firefly algorithm: recent advances and applications. *Int. J. Swarm Intelligence*, 1(1), 36–50. Retrieved from doi:10.1504/IJSI.2013.055801
- Yang, X. S., Hosseini, S. S. S., & Gandomi, A. H. (2012). Firefly Algorithm for solving non-convex economic dispatch problems with valve loading effect. *Elsevier, Applied Soft Computing*, 12(3), 1180–1186. Retrieved from doi:10.1016/j.asoc.2011.09.017
- Yao, M., Pi, D., & Cong, X. (2012). Chinese text clustering algorithm based k-means. In *2012 International Conference on Medical Physics and Biomedical Engineering (ICMPBE2012)* (Vol. 33, pp. 301–307). Elsevier. Retrieved from doi: 10.1016/j.phpro.2012.05.066, Available online at www.sciencedirect.com
- Ye, N., Gauch, S., Wang, Q., & Luong, H. (2010). An adaptive ontology based hierarchical browsing system for CiteSeerX. In *Second International Conference on Knowledge and Systems Engineering (KSE), IEEE* (pp. 203–208). Retrieved from doi: 10.1109/KSE.2010.32
- Yin, Y., Kaku, I., Tang, J., & Zhu, J. (2011). *Data Mining Concepts, Methods and Application in Management and Engineering Design*. Springer-Verlag London.
- Youssef, S. M. (2011). A New Hybrid Evolutionary-based Data Clustering Using Fuzzy Particle Swarm Optimization. In *23rd IEEE International Conference on Tools with Artificial Intelligence* (pp. 717–724). IEEE. Retrieved from doi: 10.1109/ICTAI.2011.113
- Yue, S., Wei, M., Wang, J. S., & Wang, H. (2008). A general grid-clustering approach. In *Elsevier, Pattern Recognition Letters* (Vol. 29, pp. 1372–1384). Retrieved from doi: 10.1016/j.patrec.2008.02.019
- Yujian, L., & Liye, X. (2010). Unweighted Multiple Group Method with Arithmetic Mean. In *IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA)* (pp. 830–834). Changsha: IEEE. Retrieved from doi:10.1109/BICTA.2010.5645232

- Yunrong, X., & Liangzhong, J. (2009). Water quality prediction using LS-SVM and particle swarm optimization. In *Knowledge Discovery and Data Mining, 2009. WKDD 2009. Second International Workshop on* (pp. 900–904).
- Zhang, L., & Cao, Q. (2011). A novel ant-based clustering algorithm using the kernel method. *Elsevier, Information Sciences, 181*(20), 4658–4672. Retrieved from doi:10.1016/j.ins.2010.11.005
- Zhang, L., Cao, Q., & Lee, J. (2013). A novel ant-based clustering algorithm using Renyi entropy. *Elsevier, Applied Soft Computing, 13*(5), 2643–2657. Retrieved from doi:10.1016/j.asoc.2012.11.022
- Zhang, W., Yoshida, T., Tang, X., & Wang, Q. (2010). Text clustering using frequent itemsets. *Elsevier, Knowledge-Based Systems, 23*(5), 379–388. Retrieved from doi:10.1016/j.knosys.2010.01.011
- Zhao, Y., Cao, J., Zhang, C., & Zhang, S. (2011). Enhancing grid-density based clustering for high dimensional data. *Elsevier, Journal of Systems and Software, 84*(9), 1524–1539. Retrieved from doi:10.1016/j.jss.2011.02.047
- Zhao, Y., & Karypis, G. (2001). *Criterion functions for document clustering: Experiments and analysis.*
- Zhong, J., Liu, L., & Li, Z. (2010). A novel clustering algorithm based on gravity and cluster merging. *Advanced Data Mining and Applications, 6440*, 302–309. Retrieved from doi:10.1007/978-3-642-17316-5_30
- Zhu, Y., Fung, B. C. M., Mu, D., & Li, Y. (2008). An efficient hybrid hierarchical document clustering method. In *IEEE, Fifth international conference on Fuzzy systems and knowledge discovery* (Vol. 2, pp. 395–399). Retrieved from doi:10.1109/FSKD.2008.159