ADAPTIVE FIREFLY ALGORITHM FOR HIERARCHICAL TEXT CLUSTERING

DOCTOR OF PHILOSOPHY
UNIVERSITI UTARA MALAYSIA
2016
Abstrak


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
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 ......................................... 160
  6.3.1.1 Number of Clusters between WFA_RM and WFA ......................... 160
  6.3.1.2 Performance Metrics between WFA_RM and WFA ....................... 161
  6.3.1.3 Paired Samples T-test between WFA_RM and WFA ..................... 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

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\textsubscript{R}. .......................................133
Table 5.2 Internal and relative quality metrics: WFA vs. WFA\textsubscript{R}. ........................134
Table 5.3 Average number of clusters: WFA vs. WFA\textsubscript{R}. ..................................139
Table 5.4 Summary of quality performance: WFA vs. WFA\textsubscript{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. FAK-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. FAK-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. FAK-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. FAK-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. FA K-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\textsubscript{RM} vs. PGSCM vs. DCPG using different datasets. ................................................................. 216
Table 7.17 External quality F-measure (average, best, worst, standard deviation): WFA\textsubscript{RM} vs. PGSCM vs. DCPG using different datasets. ................................................................. 218
Table 7.18 External quality Entropy (average, best, worst, standard deviation): WFA\textsubscript{RM} vs. PGSCM vs. DCPG using different datasets. ................................................................. 219
Table 7.19 Internal quality ADDC (average, best, worst, standard deviation): WFA\textsubscript{RM} vs. PGSCM vs. DCPG using different datasets. ................................................................. 221
Table 7.20 Relative quality DBI (average, best, worst, standard deviation): WFA\textsubscript{RM} vs. PGSCM vs. DCPG using different datasets. ................................................................. 222
Table 7.21 Relative quality DI (average, best DI, worst DI, standard deviation): WFA\textsubscript{RM} vs. PGSCM vs. DCPG using different datasets. ................................................................. 224
Table 7.22 Summary of quality performance results: WFA\textsubscript{RM} vs. PGSCM vs. DCPG. ........................................................................................................ 225
Table 7.23 The P-value between WFA\textsubscript{RM} & dynamic methods using average purity results (sig 2 tailed) with different datasets. ................................................................. 228
Table 7.24 The P-value between WFA\textsubscript{RM} & dynamic methods using average F-measure results (sig 2 tailed) with different datasets. ................................................................. 229
Table 7.25 The P-value between WFA\textsubscript{RM} & dynamic methods using average Entropy results (sig 2 tailed) with different datasets. ................................................................. 230
Table 7.26 The P-value between WFA\textsubscript{RM} & dynamic methods using average ADDC results (sig 2 tailed) with different datasets. ................................................................. 231
Table 7.27 The P-value between WFA\textsubscript{RM} & dynamic methods using average DBI results (sig 2 tailed) with different datasets. ................................................................. 232
Table 7.28 The P-value between WFA\textsubscript{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_R......................................................................................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 Cluster3 ........................................... 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

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACK</td>
<td>Ant Colony with Kernel method</td>
</tr>
<tr>
<td>ACO</td>
<td>Ant Colony Optimization</td>
</tr>
<tr>
<td>ACPSO</td>
<td>Automatic Clustering Particle Swarm Optimization</td>
</tr>
<tr>
<td>ALHC</td>
<td>Average Linkage Hierarchical Clustering</td>
</tr>
<tr>
<td>AP</td>
<td>Affinity Propagation</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian Information Criterion</td>
</tr>
<tr>
<td>BKM</td>
<td>Bisect K-means</td>
</tr>
<tr>
<td>C-bat</td>
<td>Bat algorithm with K-means</td>
</tr>
<tr>
<td>C-cuckoo</td>
<td>Cuckoo algorithm with K-means</td>
</tr>
<tr>
<td>C-firefly</td>
<td>Firefly algorithm with K-means</td>
</tr>
<tr>
<td>CFWS</td>
<td>Clustering based on Frequent Word Sequence</td>
</tr>
<tr>
<td>CLHC</td>
<td>Complete Linkage Hierarchical Clustering</td>
</tr>
<tr>
<td>CLIQUE</td>
<td>Clustering In QUEst</td>
</tr>
<tr>
<td>CMS</td>
<td>Clustering based on Maximal Frequent Sequence</td>
</tr>
<tr>
<td>CPSO</td>
<td>Particle Swarm Optimization with K-means</td>
</tr>
<tr>
<td>CRC</td>
<td>Corrected Rand Coefficient</td>
</tr>
<tr>
<td>C-wolf</td>
<td>Wolf algorithm with K-means</td>
</tr>
<tr>
<td>DBI</td>
<td>Davies Bouldin Index</td>
</tr>
<tr>
<td>DBSCAN</td>
<td>Density-Based Spatial Clustering of Application with Noise</td>
</tr>
<tr>
<td>DCGA</td>
<td>Dynamic Clustering Genetic Algorithm</td>
</tr>
<tr>
<td>DCPG</td>
<td>Dynamic Clustering Particle Swarm Optimization with Genetic Algorithm</td>
</tr>
<tr>
<td>DCPSO</td>
<td>Dynamic Clustering using Particle Swarm Optimization</td>
</tr>
<tr>
<td>DF</td>
<td>Document Frequency</td>
</tr>
<tr>
<td>DHC</td>
<td>Dynamic Hierarchical Compact</td>
</tr>
<tr>
<td>DHS</td>
<td>Dynamic Hierarchical Star</td>
</tr>
<tr>
<td>DI</td>
<td>Dunn Index</td>
</tr>
<tr>
<td>ES</td>
<td>Evolution Strategy</td>
</tr>
<tr>
<td>FA</td>
<td>Firefly Algorithm</td>
</tr>
<tr>
<td>FIHC</td>
<td>Frequent Itemset based Hierarchical Clustering</td>
</tr>
<tr>
<td>FTC</td>
<td>Frequent Term based Clustering</td>
</tr>
<tr>
<td>GA</td>
<td>Genetic Algorithm</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>GGCA</td>
<td>General Grid Clustering Approach</td>
</tr>
<tr>
<td>GSA</td>
<td>Gravitational Search Algorithm</td>
</tr>
<tr>
<td>GSA-KM</td>
<td>Gravitational Search Algorithm with K-means</td>
</tr>
<tr>
<td>HBMO</td>
<td>Honey Bee Mating Optimization</td>
</tr>
<tr>
<td>HCM</td>
<td>Hierarchical Clustering Method</td>
</tr>
<tr>
<td>HS</td>
<td>Harmony Search</td>
</tr>
<tr>
<td>IDF</td>
<td>Inverse Document Frequency</td>
</tr>
<tr>
<td>KCPNO</td>
<td>K-means with Particle Swarm Optimization</td>
</tr>
<tr>
<td>KFA</td>
<td>K-means with Firefly Algorithm</td>
</tr>
<tr>
<td>KHM</td>
<td>K-Harmonic Means algorithm</td>
</tr>
<tr>
<td>KPSO</td>
<td>K-means with Particle Swarm Optimization</td>
</tr>
<tr>
<td>NMI</td>
<td>Normalized mutual information</td>
</tr>
<tr>
<td>NN</td>
<td>Neural Networks</td>
</tr>
<tr>
<td>OptiGrid</td>
<td>Optimal Grid clustering</td>
</tr>
<tr>
<td>PDDP</td>
<td>Principal Direction Divisive Partitioning</td>
</tr>
<tr>
<td>PGSCM</td>
<td>Practical General Stochastic Clustering Method</td>
</tr>
<tr>
<td>PSO</td>
<td>Particle Swarm Optimization</td>
</tr>
<tr>
<td>PSOKHM</td>
<td>Particle Swarm Optimization with K-Harmonic Means</td>
</tr>
<tr>
<td>RFA</td>
<td>Reachback Firefly Algorithm</td>
</tr>
<tr>
<td>SA</td>
<td>Simulated Annealing</td>
</tr>
<tr>
<td>SAP</td>
<td>Seed Affinity Propagation</td>
</tr>
<tr>
<td>SLHC</td>
<td>Single Linkage Hierarchical Clustering</td>
</tr>
<tr>
<td>SOM</td>
<td>Self Organizing Map</td>
</tr>
<tr>
<td>STING</td>
<td>Statistical Information Grid-based method</td>
</tr>
<tr>
<td>TC</td>
<td>Term Contribution</td>
</tr>
<tr>
<td>TFIDF</td>
<td>Term Frequency–Inverse Document Frequency</td>
</tr>
<tr>
<td>TSP</td>
<td>Travelling Salesman Problem</td>
</tr>
<tr>
<td>UPGMA</td>
<td>Un-weighted Pair Group Method with Arithmetic Mean</td>
</tr>
<tr>
<td>VI</td>
<td>Validity Index</td>
</tr>
<tr>
<td>VSM</td>
<td>Vector Space Model</td>
</tr>
<tr>
<td>WFA</td>
<td>Weight-based Firefly Algorithm</td>
</tr>
<tr>
<td>WFA&lt;sub&gt;R&lt;/sub&gt;</td>
<td>Weight-based Firefly Algorithm with relocating</td>
</tr>
<tr>
<td>WFA&lt;sub&gt;RM&lt;/sub&gt;</td>
<td>Weight-based Firefly Algorithm with relocating with merging algorithm</td>
</tr>
</tbody>
</table>
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


241


