Document Clustering Using Swarm Intelligence

Thesis submitted in partial fulfillment of the requirement for the degree of

Bachelor of Science

In

Computer Science and Engineering

Under the Supervision of

Abu Mohammad Hammad Ali

By

Md. Adeeb Rizwan

ID:09201929

To

Department of Computer Science and Engineering, BRAC University

66 Mohakhali C/A, Dhaka-1212

December 2012

Declaration

I hereby declare that the Thesis entitled "Document Clustering Using Swarm Intelligence" which is submitted by me in partial fulfillment of the requirement for the award of degree of BSc in Computer Science and Engineering to the Department of Computer Science and Engineering, BRAC University, 66 Mohakhali C/A, Dhaka-1212, comprises only my original work and due acknowledgement has been made in the text to all other material used. The results of this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Name of Students:

Md. Adeeb Rizwan

ID:09201029

Date: 12th November 2012

APPROVED BY:

Abstract

With the recent rise in electronic data, and the pressing time needs of daily life, there is scope for a good document clustering system that can divide a set of documents into similar topic clusters. A lot of different algorithms have been attempted towards this end, from statistical learning methods to neural networks. In more recent years, there has been a growing interest in collective intelligence as often displayed in nature by ants and birds. We would like to do a survey of this last field, and look for possible applications of algorithms in this area for the document clustering problem.

Acknowledgement

Firstly, I would like to thank my supervisor, Abu Mohammad Hammad Ali for guiding me through my bachelor thesis. Also, I am extremely grateful to Sarwar Alam for introducing me to the field of Swarm Intelligence.

Special thanks goes to my family, friends and all our teachers for their motivation and support.

Table of Contents

List of Figures and Tables

Chapter 3: Document Clustering

Section 3.1 Information Extracting

Chapter 4: Experimental Results and Evaluation

Section 4.1 Accuracy Measurement

Chapter 1: Introduction

1.1 Motivation

Clustering is a way of representing large number of data in a set or group. These groups represent similarities between the data. With the rise in electronic data there is a need for managing these data effectively, and how fast you can access these data also depends on how well you organize them.

Clustering:

A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A *cluster* is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. So, the goal of clustering is to determine the intrinsic grouping in a set of unlabeled data.

Clustering can be considered as the most important *unsupervised learning* problem. So, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data.

Applications of Clustering:

Clustering algorithms can be applied in many fields, for instance:

-Information system**:** document classification, data clustering;

-Marketing**:** finding groups of customers with similar behavior given a large database

-Biology**:** classification of plants and animals given their features;

-Libraries**:** book ordering;

-Insurance**:** identifying groups of motor insurance policy holders with a high average for checking for frauds;

-City-planning**:** identifying groups of houses according to their house type, value and geographical location;

-Earthquake studies**:** clustering observed earthquake epicenters to identify dangerous zones;

1.2 Report Overview

The following report consists of five chapters. Chapter 1 the Introduction highlights the motivation behind the thesis.

Chapter 2, Literature Review, outlines all information relevant to the thesis.

Chapter 3 illustrates the details of this thesis experiment.

In Chapter 4 I've mentioned the results obtained and evaluation of that result.

Finally, Chapter 5 is a small conclusion including a comparison and suggestions about improvements and future goals.

Chapter 2: Literature Review

2.1 Swarm Intelligence

Swarm Intelligence (SI) is the property of a system whereby the collective behaviors of (unsophisticated) agents interacting locally with their environment cause coherent functional global patterns to emerge. These agents represent decentralized and self-organized system, natural or artificial. However, the inspiration often comes from nature, especially biological systems. The agents follow very simple rules, and although there is no centralized control structure dictating how individual agents should behave whether its local or to a certain degree random, interactions between such agents lead to the emergence of "intelligent" global behavior, unknown to the individual agents. Natural examples of SI include ant colonies, bird flocking, animal herding, bacterial growth, and fish schooling.

The expression swarm intelligence was introduced by Gerardo Beni and Jing Wang in 1989, in the context of cellular robotic systems.

Eug´ene Marais (1872-1936), one of the first to study termite colonies, published his work in The Soul of the Ant. The research was originally published in Afrikaans in 1925

and was later translated in English. Marais's researches showed the organic unity of the termitary and compared it with the how human body functions as a system.

Maurice Maeterlinck (1862–1949), published The Life of the White Ant, largely drawn from Marais's articles. It was the translated and enriched version of Marais's researches.
Pierre-Paul Grass´e (1959) postulated on the mechanics of termite communication in studies of their nest construction behavior. He introduced the term "stigmergy" to refer to termite behavior. He defined it as: "Stimulation of workers by the performance they have achieved."
It basically means that an agent's actions leave signs in the environment, signs that the other agents as well as the agent itself sense and that determine and incite their subsequent actions.
Stigmergy is now one of the key concepts in the field of swarm intelligence.
Deneubourg et al. (1990) studied pheromonal communication as an example of stigmergy.
Noting that dead ants were carefully removed from an ant colony's nesting area, an entomologist strewed the area with several thousand dead ants. On the following day he observed numerous small piles of dead ants. The second day after strewing brought a smaller number of large piles. By the third day there were only one (or sometimes two) piles. He conjectured that a probabilistic explanation of this behavior as follows: Faced with a dead ant, an ant picked it up with probability inversely proportionate to the number of other dead ants in the vicinity. While carrying a dead ant, an ant put it down with

probability directly proportional to the density of dead ants in the vicinity. Thus the coordinated behavior of piling dead ants in a single pile resulted from a simple local rule controlling the behavior of a single ant. The piling behavior is accounted for without postulating communication between the ants.

From these studies, Marco Dorigo implemented the first algorithmic models of foraging behavior in 1992. The algorithm is now known as the Ant-colony optimization and is a widely used algorithm which was inspired by the behaviors of ants, and has been effective solving discrete optimization problems related to swarming.

2.1.1 Swarm as Agents

In artificial intelligence, an intelligent agent (IA) is an autonomous entity which observes through sensors and acts upon an environment using actuators and directs its activity towards achieving goals . Intelligent agents may also learn or use knowledge to achieve their goals. They may be very simple or very complex: a reflex machine such as a thermostat is an intelligent agent, as is a human being, as is a community of human beings working together towards a goal.

In swarm Intelligence each entity of a swarm is used as agent.

Swarm behavior, or swarming, is a collective behavior exhibited by animals of similar

size which aggregate together, perhaps milling about the same spot or perhaps moving *en masse* or migrating in some direction. As a term, *swarming* is applied particularly to insects, but can also be applied to any other animal that exhibits swarm behavior.

2.1.2 Aspects of Swarm Intelligence

The field of swarm intelligence is vast. There are many researches about this area. One of the most used algorithms are,

Ant colony optimization (ACO) is a class of optimization algorithms modeled on the actions of an ant colony. ACO methods are useful in problems that need to find paths to goals. Artificial 'ants'—simulation agents—locate optimal solutions by moving through a parameter space representing all possible solutions. Natural ants lay down pheromones directing each other to resources while exploring their environment. The simulated 'ants' similarly record their positions

and the quality of their solutions, so that in later simulation iterations more ants locate better solutions.

Particle swarm optimization (PSO) is a global optimization algorithm for dealing with problems in which a best solution can be represented as a point or surface in an n-dimensional space. Hypotheses are plotted in this space and seeded with an initial

velocity, as well as a communication channel between the particles. Particles then move through the solution space, and are evaluated according to some fitness criterion after each time step. Over time, particles are accelerated towards those particles within their communication grouping which have better fitness values. The main advantage of such an approach over other global minimization strategies such as simulated annealing is that the large number of members that make up the particle swarm make the technique impressively resilient to the problem of local minima.

Ant Clustering Algorithm (ACA) is another widespread used aspect of swarm intelligence. Its used for clustering data such as documents, web pages and so on. ACA is used to cluster unlabeled data. It has better accuracy then the K-means and Heuristic algorithms.

2.1.3 SI Application

Swarm Intelligence-based techniques can be used in a number of applications. The U.S. military is investigating swarm techniques for controlling unmanned vehicles. The European Space Agency is thinking about an orbital swarm for self-assembly and interferometry. NASA is investigating the use of swarm technology for planetary mapping. A 1992 paper by M. Anthony Lewis and George A. Bekey discusses the

possibility of using swarm intelligence to control nanobots within the body for the purpose of killing cancer tumors. Conversely al-Rifaie and Aber have used Stochastic Diffusion Search to help locate tumors. Swarm intelligence has also been applied for data mining.

The Ant Colony is used in many application for SI. Ant based routing is used for finding the shortest path. Ant clustering is used for clustering problems.

2.2 Related Work

There are many approaches to document clustering. The common ones being

- Hierarchical approach
- K-means clustering

The first one is the hierarchical based algorithm which includes, single link, complete linkage, group average and Ward's method.

Single-linkage: In single-link clustering or single-linkage clustering, the similarity of two clusters is the similarity of their most similar members. This single-link merge criterion is local. A drawback of this method is the so-called chaining phenomenon. Clusters may be forced together due to single elements being similar to each other, even though many

of the elements in each cluster may be very dissimilar to each other.

Complete linkage: In complete-link clustering, the similarity of two clusters is the similarity of their most dissimilar members. This is equivalent to choosing the cluster pair whose merge has the smallest diameter. This complete-link merge criterion is non-local; the entire structure of the clustering can influence merge decisions. This results in a preference for compact clusters with small diameters over long, straggly clusters, but also causes sensitivity to outliers. A single document far from the center can increase diameters of candidate merge clusters dramatically and completely change the final clustering.

Group average: Group-average clustering evaluates cluster quality based on all similarities between documents, thus avoiding the pitfalls of the single-link and complete-link criteria, which equate cluster similarity with the similarity of a single pair of documents.

Ward's method: Also known as Ward's minimum variance method, is a special case of the objective function approach originally presented by Joe H. Ward, Jr. Ward suggested a general hierarchical clustering procedure, where the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function. This objective function could be "any function that reflects the investigator's purpose."

In the Hierarchical approach by aggregating or dividing, documents can be clustered into hierarchical structure, which is suitable for browsing. However, such an algorithm usually suffers from efficiency problems.

The other algorithm is developed using the K-means algorithm and its variants.

The K-Means is a simple clustering algorithm used to divide a set of objects, based on their attributes/features, into *k* clusters, where *k* is a predefined or user-defined constant. The main idea is to define *k* centroids, one for each cluster. The centroid of a cluster is formed in such a way that it is closely related (in terms of similarity function) to all objects of that cluster.

The k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centres. Usually, it is of greater efficiency, but less accurate than the hierarchical algorithm.

The very recent approach to document clustering is inspired by nature. Using swarm intelligence this problem can be tackled with better results. We look to use ant-clustering algorithms to create an efficient document clustering mechanism.

2.3 Ant Clustering Algorithm

Ant Clustering Algorithm mainly focuses on Brood care and Cemetery organization of ant colony.

- Brood care

Larvae are sorted in such a way that different brood stages are arranged in concentric rings. Smaller larvae are located in the center, with larger larvae on the periphery.

- Cemetery Organization

Many ants cluster corpses to form cemetery. Each ant moves individually showing random behavior while picking up or deposition of corpse. The decision to pick up or drop a corpse is based on local information of the ant's current position. This behavior forms a complex clustering formation.

This approach is focused on an ant-clustering algorithm proposed by Lumer and Faieta as a development of the ideas introduced by Deneubourg et al.

Ants tend to cluster all dead bodies in specific regions of the environment, thus forming piles of dead bodies. The first ant clustering algorithm inspired by this clustering behavior of ants was introduced in "The Dynamics of Collective Sorting: Robot-Like Ant and Ant-Like Robot", J. Deneubourg, N. Goss, N. Franks, A. Sendova- Franks, C. Detrain and L. Chrétien, where a population of robots had to group together objects without any central control. Lumer and Faieta adapted the robots ant-clustering algorithm for the analysis and classification of numerical data, thus introducing the standard ant-clustering algorithm
(ACA). Since its proposal, in 1994, the ACA has passed through some modifications and has been applied to several domains, from data mining, to graph partitioning, to text-mining. Independently of the application domain and particular version of the algorithm, ant-clustering algorithms based on ACA follow a set of basic, general principles.

2.4 ACA with modified Agents

The conventional ACA algorithm works as the following:

- create an NxN grid
- randomly place object in grid
- deploy default number of agents to perform clustering

As per this algorithm all agents behaves the same way except for their random traversal path.

All agents has 2 major tasks.

- compute pickup
- compute drop

Once an agent find an object it calculates the pickup probability from stigmergy value it receives. Upon picking up an object it move randomly about the grid until its has other similar

object in the same vicinity. Once in preferred vicinity it drops the object. All agents carry the same operation. After some period clusters emerge in grid.

For this thesis I have used a modified version of the ACA .Instead of using same agents I classified them in;

- seeker
- collector

Instead of using the same agents two types of agents are used. The seeker is initially deployed.

It seeks and picks up the first object it encounters. Once an object is discovered the seeker deploys a default number of collectors. The collectors act like the conventional ACA agents however instead of comparing stigmergy from local vicinity it calculates the seeker's object's term frequency. The collector cluster once they encounter an object with similar term frequency of seekers object.

Chapter 3: Document Clustering

Document clustering using modified ACA is done in two steps.

- Information extraction
- Clustering

3.1 Information Extracting

The Information extraction is done using wvtools library. Words are extracted from document as term to calculate the term frequency.

Terms are described as,

- word e.g. “airplane”

- n-gram e.g. “airp”, “irpl”, “rpla”

- collocation e.g. “white house”

The term frequency refer to the number of times a term occurs in a document. Stop words are ignored while extracting terms.

Stop Words are described as,

- "am", "is", "are"

- "the", "he", "she"

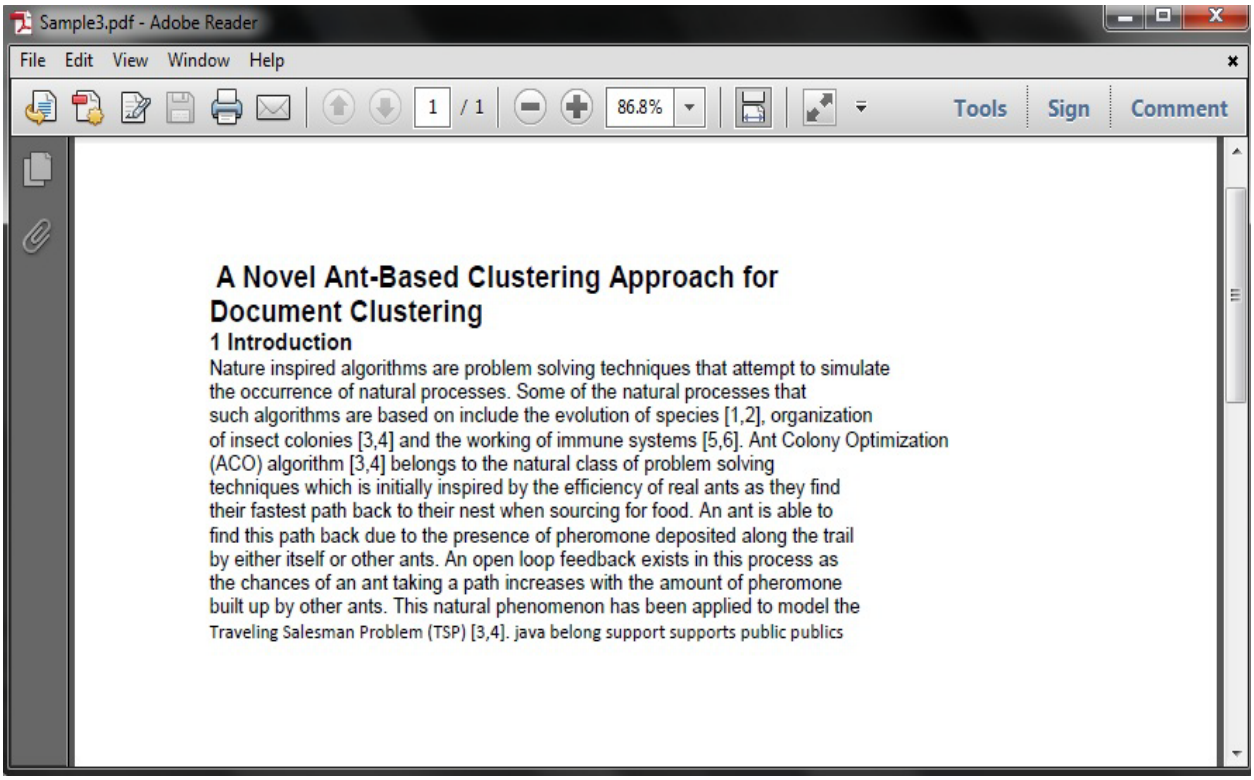

figure 3(a): a sample document

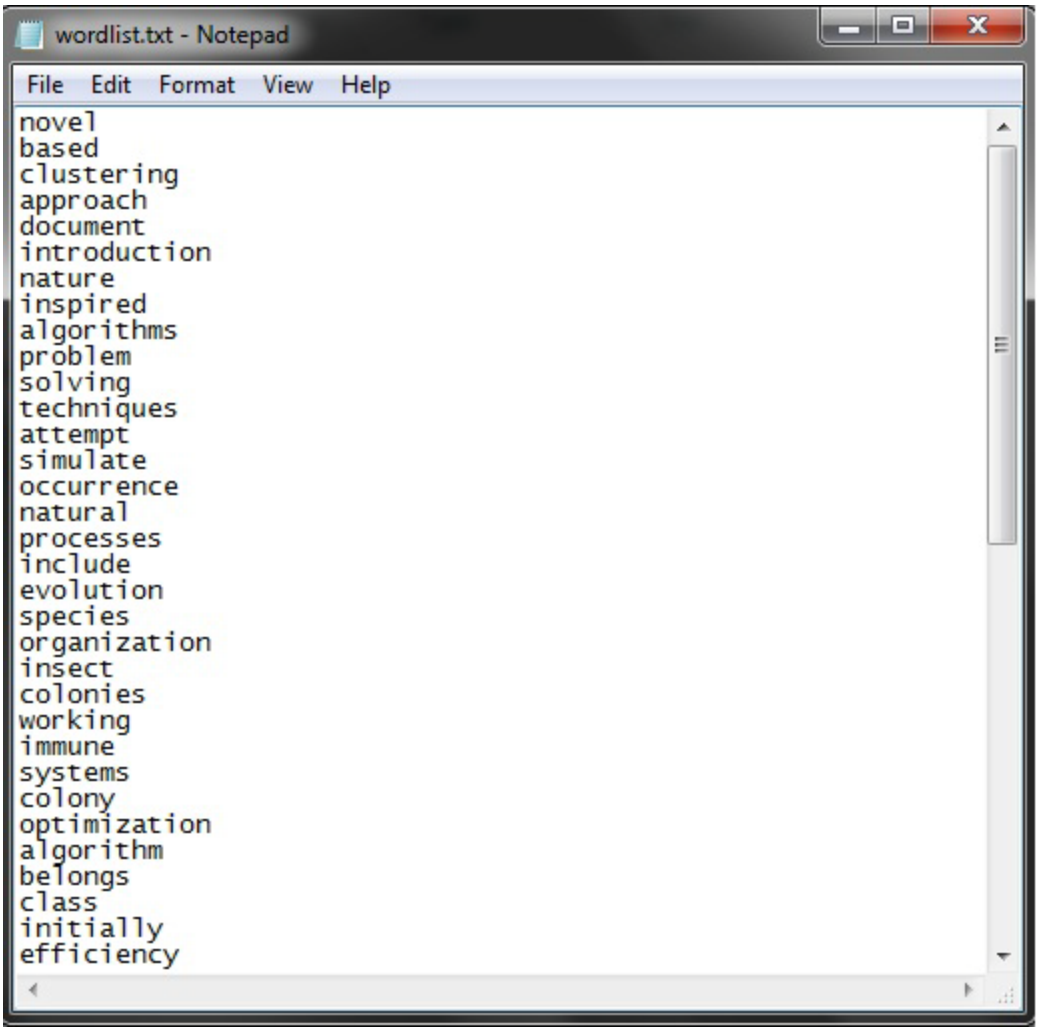

figure 3(b): non repeated wordlist

Once the extraction is complete the document's corresponding object with term frequency is created and placed in grid for clustering.

3.2 Clustering

For clustering initially the seeker is deployed. Once the seeker find an object it deploys its collectors who collect only the objects with similar term frequency of seekers object. After doing so, it clusters all the objects around the seeker. Once one seeker is occupied another seeker is deployed. Each non initial seeker avoids objects with similar term frequency of previous seekers object. the Number of seekers represent the number of clusters.

Chapter 4: Experimental Results and Evaluation

4.1 Accuracy Measurement

Accuracy measurement is done using the following terms,

True Positive(tp): True positive means number of correct result we are looking for.

True Negative(tn): Means correct absence of the irrelevant result in our system result.

False Positive(fp) : Means wrong result in our system output(Unexpected output).

False Negative(fn): Means Missing expected outputs. (The result should be included in system output but not there).

From these terms accuracy is calculated by,

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Cluster	Tp	Tn	Fp	Fn
C1	9	39	1	1
C2	8	40	0	2
C3	10	40	0	0
C4	8	38	2	2
C5	9	40	0	1

C1 cluster accuracy: 96%

C2 cluster accuracy: 96%

C3 cluster accuracy: 100%

C4 cluster accuracy: 92%

C5 cluster accuracy: 98%

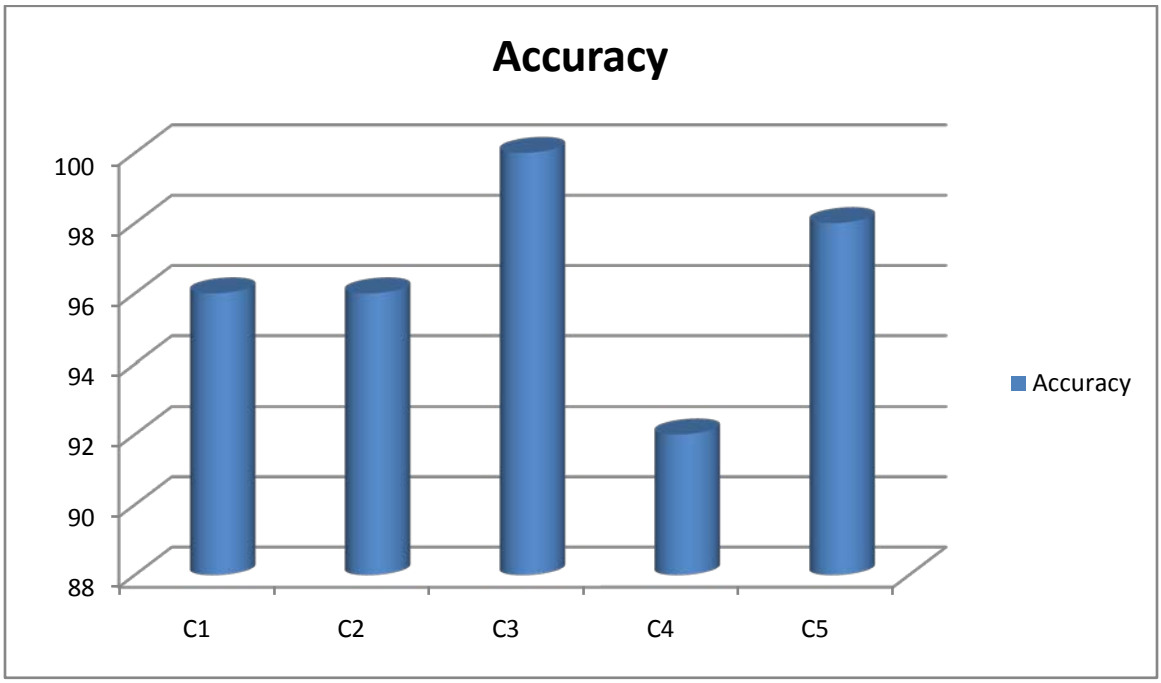

figure 4.1(a): accuracy of each clusters

Chapter 5: Conclusion and Future Work

According to research done by Yulan He, Siu Cheung Hui, and Yongxiang Sim
School of Computer Engineering, Nanyang Technological University
Nanyang Avenue, Singapore 639798, the performance of K-means and Hierarchical approach is given below:

	F-Measure			
Method	*Subset 1*	*Subset 2*	*Subset 3*	*Subset 4*
AHC	0.665	0.654	0.700	0.631
K-means	0.794	0.580	0.513	0.624

this experiment is done using 1200 unlabeled documents. This thesis has proposed a modified version of ACA with higher accuracy however the results were obtained using Training documents designed to similar as per clusters.

In my thesis I analyzed only a small part of Swarm Intelligence. There is a lot more that needs to be and can be done in this field to achieve the final goal of clustering a large number of data set using real world documents.

The Future work will be to implement this algorithm to perform for real unlabeled and to see how the algorithm fares with existing ones.

BIBLIOGRAPHY

[1] J. L. Deneubourg, S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, and L. Chretien. 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*, pages 356–363, Cambridge, MA, USA, 1990. MIT Press.

[2] Lumer E. D. and Faieta B. Diversity and adaptation in populations of clustering ants. In Cli D., Husbands P., Meyer J., and Wilson S., editors, *Proceedings of the Third International Conference on Simulation of Adaptive Behaviour: From Animals to Animats 3*, pages 501–508, Cambridge, MA, 1994. MIT Press.

[3] B. Wu, Y. Zheng, S. Liu, and Z. Shi. Csim: a document clustering algorithm based on swarm intelligence. In *Proceedings of the 2002 congress on Evolutionary Computation*, Honolulu, USA, 2002

[4] G. Beni, "The Concept of Cellular Robotic Systems", *Proc. of the IEEE Int. Symp. on Intelligent Control*, pp. 57-62, (1988).

[5] G. Beni, and J. Wang, "Swarm Intelligence", *Proc. of the 7th Annual Meeting of the Robotics Society of Japan*, pp. 425-428, (1989).

[6] Swarm Intelligence: http://www.sce.carleton.ca/netmanage/tony/swarm.html

[7] E. Lumer, and B. Faieta, "Exploratory Database Analysis via Self-Organization", *Unpublished Manuscript*, (1995).

[8] VSM: http://en.wikipedia.org/wiki/Vector_space_model

[9] Marco Dorigo : http://iridia.ulb.ac.be/~mdorigo/HomePageDorigo/

[10] Accuracy and precision: http://en.wikipedia.org/wiki/Accuracy_and_precision