Web-Page Clustering Using Swarm Intelligence

By

Nikhil Brahmwar

12MCEC44

CSE Department

INSTITUTE OF TECHNOLOGY,NIRMA UNIVERSITY

AHMEDABAD

MAY 2014

Web-Page Clustering Using Swarm Intelligence

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

M.TECH in Computer Science and Engineering

By

Nikhil Brahmwar

12MCEC44

Guided By

Prof. Sapan H. Mankad

CSE Department

INSTITUTE OF TECHNOLOGY,Nirma University

AHMEDABAD

May, 2014

Certificate

This is to certify that the Major Project Report entitled “” Web-Page Clustering Using Swarm Intelligence” submitted by Nikhil Brahmwar (Roll No: 12MCEC44), towards the partial fulfillment of the requirements for the degree of M.Tech in Computer Science and Engineering of Nirma University, Ahmedabad is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this major project, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Prof. Sapan H. Mankad
Guide & Assistant Professor,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Prof. Vijay Ukani
Associate Professor
Coordinator M.Tech - CSE
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Sanjay Garg
Professor and Head,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr K Kotecha
Director,
Institute of Technology,
Nirma University,
Ahmedabad

Declaration

I hereby informs that the Major Project entitled Web-Page Clustering Using Swarm Intelligence contains my orignal work towards the partial fulfillment of requirements for the degree of M.tech in Computer Science and Engineering at Nirma University and has not been submitted elsewhere for a degree or any other purpose.Due acknowledgement has been made in the content to all other material used

Nikhil Brahmwar

Supported by

Prof. Sapan H. Mankad

Acknowledgements

I express my thanks to my Guide Prof. Sapan H.Mankad, CSE Depatment,Nirma University for his guidence and support during the entire project work.I would also like to thank him to encourage me when I stuck during my Project Work.His Guidence has helped me to face the problems and ways to come out from that.

I would also like to thank all the faculty members for their suggestions for Project, Cooperation and facility provided to us during Project work.And lastly I would like to thanks my friend and family to motivate me and their belief in me.

- Nikhil Brahmwar

12MCEC44

Abstract

Clustering is unsupervised approach.Clustering is a procedure which segments a given information set into homogeneous group focused around given characteristics such that comparable items are kept in a same group while disparate articles are in different groups.Now a days large amount of information is available on Internet so Web mining is solution for gaining Knowledge and information.

In this dissertation, we use web-page clustering.It is used to cluster the search results containing same type of Documents.This will improve the similiarity search results which cointain similiar type of content.

In this dessertation LSI(Latent Semantic Indexing) is used for feature Reduction and then Hybrid PSO(particle Swarm Optimization) Technique is applied on the dataset to improve the result of Web page Clustering.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

It is well realized that the internet may be recognized as an enormous and worldwide data focus. A site typically holds extraordinary measures of data circulated through many pages. Without legitimate direction, a guest frequently meanders capriciously without going by imperative pages, loses investment and leaves the site sooner than anticipated. This thought is at the premise of the extraordinary enthusiasm about web data mining both in the scholarly and the modern world.

The term Web mining indicate the utilization of information mining procedures to naturally uncover Web archives and administrations, extricate data from Web assets, and reveal general examples on the Web. Through the years, Web mining examination has been reached out to blanket the utilization of information mining and comparative methods to uncover assets, examples, and learning from the Web and Web related information. Characterized as, "the finding and dissection of helpful data from the Internet", web mining experience substantial volumes of site page information rapidly and endeavor to change that information in a manner that will upgrade the revelation of new learning. Mining uses a consolidation of numerous controls: from detail and example distinguishment to manmade brainpower and machine taking in.

Despite the fact that web mining has been effective in a number of its requisitions, it is still a moderately new field. New systems and developments are discovered often, but there is much potential for development.

1.1 Scope of the Work

To achieve the objective of developing a Particle Swarm Optimization for effective clustering, the work seeks to develop an optimal number of clusters using Particle Swarm Optimization. The parameter for optimization is the accuracy and additionally, the number of clusters so as to address the important issue of ease in comprehension. The approach considers designing a Hybrid algorithm .Consequently the degree is restricted to proposing another grouping calculation that is Consolidating particle Swarm Advancement and K-means Clustering.

Chapter 2

Web-Page Clustering

Site page Clustering is one of the significant and most essential preprocessing steps in web mining examination. In this things to be contemplated are pages. Site page bunching assembles pages in gatherings, in light of similitude or other relationship measures.

2.1 Clustering

Grouping is an information mining method used to place information components into related gathering without development learning of the gathering definitions.
Clustering is a procedure which segments a given information set into homogeneous gatherings focused around given characteristics such that comparable items are kept in a gathering while disparate articles are in diverse gatherings. It is the most imperative unsupervised taking in issue. It manages discovering structure in an accumulation of unlabeled information.

Clustering is the procedure of gathering the information into classes or bunches with the goal that protests inside a bunch have high closeness in relationship to an alternate, yet are extremely unlike questions in different groups. Information bunching is under enthusiastic advancement and is connected to numerous requisition zones including business, science, prescription, science, and so on. Owing to the gigantic measures of information gathered in databases, group investigation has as of late turned into an exceedingly dynamic subject in information mining exploration. For group investigation

to work effectively and successfully, as numerous writings have exhibited, there are the accompanying common necessities of bunching in information mining

1. **Versatile:** Numerous bunching calculations work well on little information sets holding fewer than a few hundred information objects; then again, a substantial database may hold a huge number of articles. Grouping on a specimen of a given huge information set may prompt predispositioned results. Profoundly versatile bunching calculations are required.

2. **Capability to manage diverse sorts of characteristics:**Numerous calculations are intended to bunch interim based (numerical) information. Nonetheless, provisions may oblige bunching different sorts of information, for example, parallel, straight out (ostensible), and ordinal information, or mixtures of these information sorts.

3. **Revelation of bunches with subjective shape:**Numerous bunching calculations focus groups focused around Euclidean separation measures. Calculations focused around such separation measures have a tendency to discover round groups with comparative size and thickness. In any case, a group could be of any shape. It is essential to create calculations that can catch bunches of subjective shape.

4. **Insignificant necessities for space information to focus data parameters:**Numerous bunching calculations oblige clients to include certain parameters in bunch dissection, (for example, the amount of sought bunches). The grouping results could be very touchy to enter parameters. Parameters are frequently troublesome to focus, particularly for information sets holding high-dimensional items. This loads clients, as well as makes the nature of grouping troublesome to control.

5. **Capacity to manage noisy information:**Most true databases hold outliers or missing, obscure, or incorrect information. Some grouping calculations are delicate to such information and may prompt bunches of low quality.

6. **Cold-heartedness to the request of information records:**Some grouping calculations can't join recently embedded information into existing bunching structures and, rather, must focus another bunching starting with no outside help. Some bunching calculations are delicate to the request of information. That is, given a

set of information items, such a calculation may return significantly distinctive clusterings relying upon the request of presentation of the data objects. It is paramount to create incremental bunching calculations and calculations that are heartless to the request of data.

7. **Large Axes:**A database or an information warehouse can hold a few measurements or characteristics. Numerous grouping calculations are great at taking care of low-dimensional information, including just two to three measurements. Human eyes are great at judging the nature of bunching for up to three measurements. Discovering groups of information questions in high dimensional space is testing, particularly acknowledging that such information could be scanty and profoundly skewed.

2.2 Web Page Clustering

A grouping [15] framework could be valuable in web scan for gathering query items into nearly related sets of records. Bunching can enhance closeness seek by concentrating on sets of important records. Bunching is likewise a profitable procedure for dissecting the Web. Matching the substance based grouping and the hyperlink structure can uncover examples, duplications, and other intriguing structures on the Web.

2.2.1 Representation of website page:

Grouping might be connected to any set of articles as long as a suitable representation of these items exists. The most widely recognized representation, which likewise works for other machine taking in and information mining routines, (for example, grouping), is the characteristic esteem (or characteristic quality) representation. In this representation various (characteristics) are distinguished for the whole populace, and each one article is spoken to by a situated of property estimation sets. On the other hand, if the request of the characteristics is altered, a vector of qualities (information focuses) could be utilized. The report vector space model is precisely the same kind of representation, where the characteristics are terms.

Vector Space Model

It characterizes reports [15] as points in a multiaxes Euclidean space where the axes (measurements) are spoken to by terms. Contingent upon the kind of vector parts, there are three fundamental renditions of this representation: Boolean, term recurrence, and term recurrence reverse archive recurrence.

As we said prior, the Boolean representation is straightforward, simple to register, and works well for report grouping and bunching. Then again, it is not suitable for magic word seek on the grounds that it doesn't permit record positioning. Subsequently, we center here on the TFIDF representation.

In the term recurrence approach, the directions of the record vector are spoken to as a capacity of the term numbers, typically standardized with the report length. For each one term and each one archive , the TF measure is registered. This is possible in diverse ways.

- The aggregate number of terms in the record:

$$tf(a_j, bk) = \begin{cases} 0 & if \ \ s_{pq} = 0 \\ \frac{s_{pq}}{\sum_{l=1}^{m} s_{lq}} & if \ \ s_{pq} > 0 \end{cases} \tag{2.1}$$

- Utilizing the most extreme of the term include over all terms the report:

$$tf(a_j, bk) = \begin{cases} 0 & if \ \ s_{pq} = 0 \\ \frac{s_{pq}}{max_k s_{lq}} & if \ \ s_{pq} > 0 \end{cases} \tag{2.2}$$

In the boolean and tf representations, each one direction of a record vector is figured by regional standards, considering just the specific term and archive. This implies that all tomahawks are recognized to be similarly imperative. Be that as it may, terms that happen much of the time in reports may not be identified with the substance of the archive. This is thusly expands the measure of the ensuing set and makes record positioning troublesome if this term is utilized within the question. The same impact is created by

stopwords, for example, is,of,the,or,as,to and at and is one motivation to dispose of them.

The essential thought of the converse report recurrence methodology is to scale down the directions for a few tomahawks, comparing to terms that happen in numerous documents.for each one term ti the IDF measure is processed as an extent of archives where happens regarding the aggregate number of records in the gathering

The sorts of parameters focus the sort of archive representation:

- The least difficult approach to utilize a term as a characteristic in a report representation is to check whether the term happens in the record. Hence, the term is acknowledged as a Boolean property, so the representation is called Boolean.

- The worth of a term as a characteristic in a report representation may be the amount of events of the term (term recurrence) in the record or in the whole corpus. Report representation that incorporates the term frequencies yet not the term positions is known as a pack of-words representation on the grounds that formally it is a multiset or sack (a sort of set in which every thing may happen various times).

- Term positions may be incorporated alongside the recurrence. This is a complete representation that jam the majority of the data and may be utilized to produce the first report from its representation.

Pertinence Positioning

- The Boolean catchphrase pursuit is basic and productive, yet it gives back a set (unordered accumulation) of archives (can't exceptionally distinguish the ensuing records).

- The result is to rank reports in the reaction set by pertinence to the question and present to the client a requested rundown with the top-positioning archives first.

- The Boolean term archive lattice can't, on the other hand, give requesting inside the archives matching the set of pivotal words.

2.2.2 Records Representation:

- Records are tokenized; that is, all punctuation imprints are uprooted and the character strings without spaces are acknowledged as tokens .
- All characters in the archives and in the question are changed over to upper or easier case.
- Words are decreased to their standard structure . Case in point, variant structures, for example, is and are supplanted with be, different endings are evacuated, or the words are changed into their root structure, for example, projects and programming into project. This methodology, called stemming, utilization morphological data to permit matching distinctive variants of words.
- Articles and other regular words that seem oftentimes in content archives yet don't bring any importance or help recognize records are called stopwords. Cases are is,on,at,to,the. These words are generally evacuated.

Chapter 3

Literature Survey

K. Premalatha [5], have describe a A New Approach for Clustering. Data Clustering is a approach of finding groups of pattern. Clustering is method that is used to group same type of data. Here various Clustering tecniques are explained such as hierarchical and optimization etc. Particle Swarm Optimization tecnique is used for clustering.one paricle is one possible solution. In general particle swarm optimization there is localbest and global besr solution. But sometime it gives local optimal solution so early even if it is not achived yet. This is one of the problem in particle swarm optimization that it gives local optima which is not correct result. In particle swarm optimization stopping condition are there when to stop such as no of iteration or if it gives same results many times.

In this paper a new tecnique is proposed to remove the Problem of local optima in general pso.This algorithm is applied on many dataset and its give good results and promising results and increase its performances.

Qinghai Bai [7],presents the investigation of Molecule Swarm Improvement. Molecule swarm enhancement is a heuristic worldwide advancement technique and additionally a streamlining calculation, which is in view of swarm discernment. It originates from the examination on the flying creature and fish group development behavior.this alogo rithms is generally utilized as a result of its simple execution and great results.this paper gives some alteration in calculation

Molecule swam streamlining is a improvement technique focused around swarm in-

telligence.the exploration of PSO is chiefly focused on the essential hypothesis of the Alogorithm,topology of the swarm.

We can combine this algoritms with other algorthm to get the good results.by using advantages of both the algorithms.

Shafiq Alam, Gillian Dobbie [8] presents a new Algorithm Evolutionary pso which is used for clustering.There are various changes done in orignal pso to improve the result of clustering.

According to this algo first we initialize the population of particle in the search space.here population is also know as swarm.In these algo a new concept of generation of population is intoduced.Particle have to manage after every generation of population for their best position in search space.

This new algo is compared with various othe Clustering Algosand orignal pso.the result shows that this new technique is more efficient than othe algorithm and gives very good results and improved results.

The main idea of this algo is the concept of generation which improves the results.The large population is important in covering complete search space.But large population may cause problem later on.So some particle remove automatically if they do not meet condition of objective function.The process take place after every generation so at last it gives optimal results.

Xiaohui Cui, Thomas E. Potok [6] presents a new technique for clustering of documents.Today if we want to gain knowledge then internet is one of the main source.we can get important infor mation which we want to know.Good Clustering algo plays an important role for organizing this information.

Partition Algo plays an important role for partition huge data in groups.there are various well-known partition algorithms such as k-meaans.k-means is very useful.we can

easily execute it and it also gives very prominent results.It is very fast in Converging to local optima.when dimensions is very high it might not give very good results.

In this paper combition of two algorithm is used on various dataset.first particle swarm optimization is implemented then its result is given as input to k-means .This technique is known as hybrid technique.By comparing this results with these algo when they were execute alone give good results.

Chapter 4

The Proposed Algorithm

Swarm Intelligence is a counterfeit consciousness procedure based around the investigation of aggregate conduct in decentralized, self-sorted out frameworks. SI frameworks are regularly made up of a populace of straightforward executors interfacing generally with each one in turn and with nature's turf. Samples, flying creature running, fish educating, and so on...

SI is concerned with the investigation of frameworks containing numerous agents that:
◇ are generally basic.
◇ coordinate through circulated control
◇ trace their surroundings generally
◇ can correspond specifically just with their neighbors
◇ can convey in a roundabout way.
◇ misuse positive and negative sentiment instruments

4.1 Preprocessing Steps

Info an Information record in ARFF position

- Change over the web records into content archives. Case in point, utilize the "Recovery As . . . " choice of the Firefox or chrome with "Recovery as sort: Content Document (*.txt)."

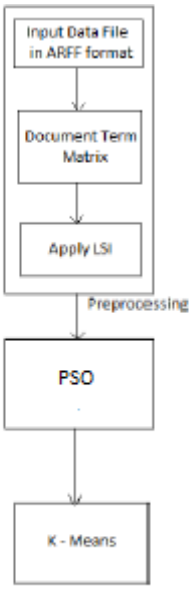

Figure 4.1: Flow of Proposed Algorithm

- Spare all archives in a solitary document . Change over it to content arrangement and analyze its substance.
- Utilize the linking of the web reports and make a content record where each one report is spoken to on a different line in plain content organization.
- Encase the archive content in quotes (") and include the record name at the start of each one line and a document header at the start of the document.

- This representation utilizes two characteristics: reportname and documentcontent, both of sort string.

Document Term Matrix

- Load the document in the Weka framework utilizing the "Open record" catch in "Preprocess" mode.
- After effective stacking the framework demonstrates to a few facts about the amount of characteristics, their sort, and the amount of cases .

- Pick the Stringtonominal channel and apply it to the first characteristic, report name.then pick the Stringtowordvector channel and apply it with "outputwordcounts= genuine".

- Now you have a record term grid stacked in Weka. Utilize the "Alter" alternative to see it in an even arrangement, where you can additionally transform its substance or duplicate it to different requisitions . Once made in Weka the table could be put away in an ARFF record through the "Spare" alternative.

LatentSemanticIndexing (LSI) Latent semantic indexing (LSI) is an indexing and recovery system that uses a numerical procedure called singular value decomposition (SVD) to recognize designs in the connections between the terms and ideas held in an unstructured gathering of content. LSI is focused around the standard that words that are utilized as a part of the same settings have a tendency to have comparable implications. A key characteristic of LSI is its capability to concentrate the calculated substance of an assemblage of content by making cooperations between those terms that happen in comparative connections.

LSI is additionally a provision of correspondence dissection, a multivariate measurable strategy, to a possibility table manufactured from word numbers in records.

4.2 Particle Swarm Optimization

PSO was motivated by the social conduct of a fledgling group. In the PSO calculation, the flying creatures in a herd are typically spoken to as particles. These particles could be recognized as straightforward operators "flying" through an issue space. A molecule's area in the multi-dimensional issue space speaks to one answer for the issue. At the point when a molecule moves to another area, an alternate issue result is generated.:this result is assessed by a wellness work that gives a quantitative worth of the result's utility. The speed and bearing of every molecule moving along each one measurement of the issue space will be modified with every era of development. In mix, the molecule's close to home encounter, Pid and its neighbors' experience, Pgd impact the development of every molecule through an issue space. The irregular qualities, rand1 and rand2, are utilized for

the purpose of fulfillment, that is, to verify that particles investigate wide inquiry space before merging around the ideal result. The qualities of constant1 and constant2 control the weight parity of pid and pgd in choosing the molecule's next development speed. For each era, the molecule's new area is processed by including the molecule's present speed, V-vector, to its area, Xvector. Numerically, given a multi-dimensional issue space, the ith molecule transforms its speed and area as indicated by the accompanying comparisons

$$Vel_k(j+1) = WVel_k(j) + Cons1[P(b) - Pok(j)] + Cons2[G(b) - Pok(j)] \tag{4.1}$$

$$Po_k(j+1) = X_k(j) + V_k(j+1) \tag{4.2}$$

where w indicates the inactivity weight component; pid is the area of the molecule that encounters the best wellness esteem; pgd is the area of the particles that encounter a worldwide best wellness esteem; constant1 and constant2 are constants and are known as increasing speed coefficients; d means the measurement of the issue space; random1, random2 are irregular values in the extent of (0, 1). The dormancy weight variable w gives the fundamental differing qualities to the swarm by changing the energy of particles to maintain a strategic distance from the stagnation of particles at the neighborhood optima.

Create and initialize an n_x-dimensional swarm;
repeat
    for *each particle* $i = 1, \ldots, n_s$ **do**
        //set the personal best position
        if $f(\mathbf{x}_i) < f(\mathbf{y}_i)$ **then**
            $\mathbf{y}_i = \mathbf{x}_i$;
        end
        //set the global best position **if** $f(\mathbf{y}_i) < f(\hat{\mathbf{y}})$ **then**
            $\hat{\mathbf{y}} = \mathbf{y}_i$;
        end
    end
    for *each particle* $i = 1, \ldots, n_s$ **do**
        update the velocity using equation
        update the position using equation
    end
until *stopping condition is true*;

Algorithm Particle Swarm Optimization

Chapter 5

Impementation Results

The results of the propose work in these thesis is given below.In these chapter various information such as technology used implementation details is also presented.

5.1 Technology

Tecnology plays an important role in any implementation.Here matlab is used for the implementation of the algorithm.Before implementation of algorithm preprocessing is nacessary .Preprocessing of dataset is done in Weka.
There are various advantages of Matlab:

- It is very user friendly .and eay to work on.
- There are error solving facilities are available.
- GUI is also one of the main feature for using Matlab.
- Coding in Matlab is easy to write and understanding.
- Matlab is very fast in execution when implementation is small.
- Large no of function already available in matlab which plays important role in reducing the size of code..
- It is used to implement in varous domain such as Datamiming,image processing etc.

Preprocessing Tool

Preprocessing is an imperative period of each implementation.after this stage its is assured to apply the algorithm.weka instrument is utilized for the Preprocessing of Dataset.It assumes an imperative part to preprocess in datamining.there are different calculations of clustering,classifcation is already executed in Weka.user can get advantage of this and can utilize its for their calculations.

5.2 Performance Measure

The following terms are used for the performance measure quantity.

- When the webpages is truly predicted to the cluster to which it belongs is called TruePositive
- When the webpages is falsely pridected to the cluster to which it doesnt belongs is called TrueNegative.
- When the webpages is truly pridected that it doesnt belongs to particular cluster then it is called FalsePositive.
- When the webpages is falsely predicted that it doesnt belongs to particular cluster then it is called FalseNegative.

The easy approach to see the structure of the mistake is to incorporate the amount of records falling in each of the classes above in a lattice called a disarray grid (likewise, possibility table) as follow:

	AssumedClusters	
ActualClasses	Pos	Neg
Positive	TPos	FNeg
Negative	FPos	TNeg

Table 5.1: Confusion Matrix

A perplexity grid might be fabricated with more than two classes and groups by utilizing more lines (for the classes) and more segments (for the bunches). In this manner,

we can characterize a summed up disarray framework for m classes and k bunches as demonstrated in Table. The number d_{ba} in each one cell shows the amount of reports from group a that have a place with class b. Presently we can characterize review and accuracy concerning class b and bunch a as takes after:

	Clusters				
Categories	1	...	a	...	d
1	d_{11}	...	d_{1a}	...	d_{1d}
.					
.					
.					
b	d_{b1}	...	d_{ba}	...	d_{bd}
.					
.					
.					
c	d_{c1}	...	d_{cj}	...	d_{cd}

Table 5.2: Confusion Matrix for c classes and d clusters

The F-measure gives a more exact record to the blunder than does the in general precision. The demonstration is that the F-measure is really the consonant mean of exactness and review. The consonant mean is utilized for averaging rates. The accuracy and review might be seen as rates, in spite of the fact that not of the same kind, yet it appears that the consonant.

5.3 Data Sets

To check the adequacy of the calculation Algo in this paper, we led investigates datasets from genuine areas. The point behind utilizing datasets for experimentation is to demonstrate the consistency of the proposed calculation in areas. The details about these dataset are listed in Table 5.3. The no of attributes in these dataset is very large, checking the execution of proposed calculation on dataset given below.

	Documents	Attributes	Clusters
syskillwebert	414	8429	4

Table 5.3: Data Sets

5.4 Experimental Results

In preprocessing, make term document matrix using StringToNominal filter and StringToWordVector filter. In StringToWordVector filter, use TF-IDF document representation, set OutputWordCount True, set LovinsStemmer as Stemmer, set useStoplist True.

Fig 5.1 shows result of preProcessing of Dataset.

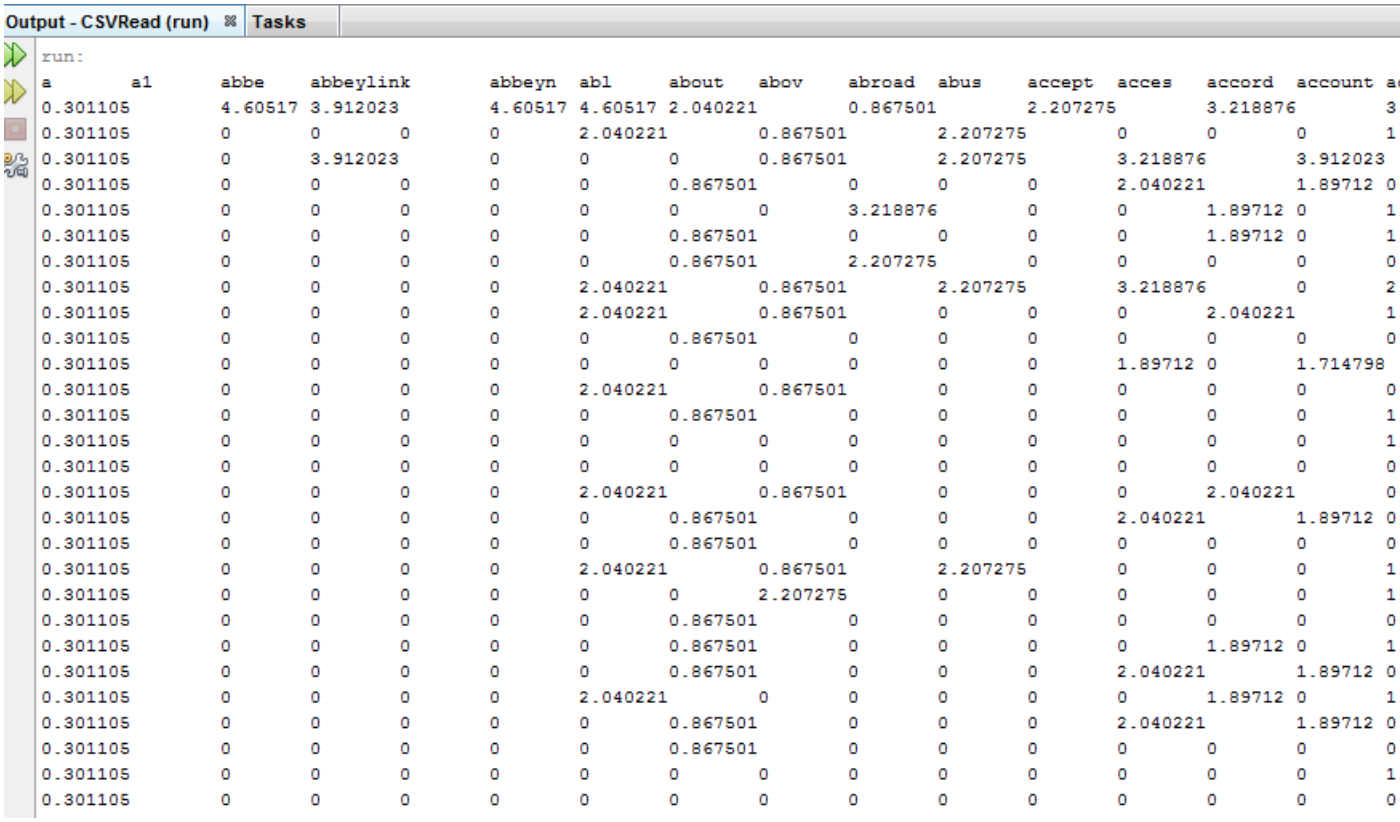

Figure 5.1: Data Sets

Above table shows the result after preprocessing of dataset.

In this phase, Some initial runs were performed and The results showing F-measure for Data Sets for two distance measures euclidean and cosine is shown in Fig

No of	**LSI**	
Attributes	**Euclidean**	**Cosine**
300	0.86588	0.64267
400	0.88277	0.67594
500	**0.88635**	0.68105
600	0.88394	0.69657

Table 5.4: F-measures for Data Set

From above table we conclude the Best result in Euclidean distance, No. of attributes 500 in LSI feature selection for Data is 0.88635.

Chapter 6

Conclusion and Future Work

In this thesis, algorithm inspired by Particle Swarm Optimization is used to cluster web pages. Web pages are preprocessed and represented in low dimensional space using Latent Semantic Indexing.

Particle Swarm optimization is first applied .The result of Particle Swarm optimization is given as input to K-means Clustering to achive the final clustering of Web-Page.

The proposed Algorithm will be tested on dataset.Other modifications will also be used to cluster the web pages.

References

[1] Bing Liu, Web DataMining: Exploring Hyperlinks, Content, and Usage Data

[2] Jiawei Han and Micheline Kamber,Data Mining: Concepts and Techniques

[3] Andries P. Engelbrecht,Computational Intelligence

[4] Rui Xu and Donald Wunsch II,Survey of Clustering Algorithms - IEEE Trans. Neural Networks,May 2005

[5] K.Premalatha,A.M Natrajan ,A New Approach for Data Clustering Based on PSO,November 2008,Vol.1.No4

[6] Xiaohui Cui and Thomas E.Potok,Document Clustering Analysis Based on Hybrid PSO ,April 2010

[7] Qinghai Bai,Analysis of Particle Swarm Optimization Algorithm - Computer and Information Science,Vol 3,No1,February 2010

[8] Shafiq Alam and Gillian Dobbie,An Evolutionary Particle Swarm Optimization Algorithm for Data Clustering - IEEE 2008

[9] Pritesh Vora and Bhavesh Oza,A survey on K-means Culstering and Particle Swarm Optimization - IJISME,VOLUME-1,Issue-3,February 2013

[10] Young-Bin Shin and Eisuke Kita,Solving two-dimensional packing problem using PSO,Feb 2012.

[11] Magnus Erik Hvass Pedersen,Good Parameters for Particle Swarm Optimization,May 2010.

[12] Goncalo Pereira, Peter Andreae and Xiaoying Gao, "Particle Swarm Optimization",April 15,2011.

[13] Ching-Yi, Fun Yei-Particle Swarm OPtimization Clustering Algorithm-Intternattiionall Journall of Ellecttriicall Engiineeriing.. voll.. 13-2006

[14] Eric Bonabeau,Marco Dorigo,Guy Theraulaz - Swarm Intelligence,Oxford University Press 1999.

[15] Z. Markov and D. T. Larose, DATA MINING THE WEB Uncovering Patterns in Web Content, Structure, and Usage. John Wiley and Sons, 2007.