ISSN:0975-9646

Evaluation of Document Clustering Approach based on Particle Swarm Optimization Method for Twitter Dataset

Baljeet Kaur
M.TECH CSE, RIEIT Ropar, Punjab.

Sheetal Kundra
Asstt. Professor, Dept. of CSE & IT, Rayat Institute of Engg. & IT, Ropar.

Harish Kundra
HOD CSE & IT, Rayat Institute of Engg. & IT, Ropar.

Abstract-This research present a survey on clustering of twitter dataset with the help of particle swarm optimization technique. Twitter dataset having no of tweets which may include video, photos and 140 character of text and link etc. Recently tweets achieved alot of significance because most of social website like facebook and twitter permit users to post short message on their frontpage and **their capability to spread information rapidly . Particle swarm optimization (PSO) technique is the best method to solve clustering problem. PSO optimizes problem with large no of candidate solutions. PSO is a procedure which solve a problem by iteratively updating a candidate solution .In PSO algorithm problem is optimized by no of candidate solutions .Here different particles are used for finding the appropriate solution where every particle have some velocity and position .**

Keywords: Document Clustering, PSO, Swarms ,Twitter, inertia component, cognitive component, acceleration coefficient .

INTRODUCTION

Document clustering is process of recognition of almost identical classes of objects [7]. A cluster is a collection of objects which are identical .It is the subset of objects in which the distance between two objects in the cluster is less .Quality of good clustering method based upon the values of similarity .Principle of document clustering based upon the fact that the intra-cluster similarity should be high and inter-cluster similarity should be low. A clustering method that will produce high quality clusters is called good cluster method.Various types of applications like search engines and document browsers used document clustering as an efficient tool for searching .This clustering tool gives good and complete outlook of the document's information .Specific problem in text clustering where the various methods of document clustering works: they are document's high dimensionality ,dataset of huge size , ability to understand the cluster characterization .There is also high demand of hierarchical document clustering so that document should be searched fast which is based upon the user's need and subjects of increasing specificity.PSO is a procedure which solve a problem by iteratively updating a candidate solution .In PSO algorithm problem is optimized by no of candidate solutions .Here different particles are used for finding the appropriate solution where every particle have some velocity and position .To find the movement of particles in the search space there is some specific mathematical formulae to find the velocity and position of particles .Particles local best known position influence each particles movement but, in the search-space it is also guided toward the best known positions, when better positions are found by other particles then this best position of particle is updated . By updating the position swarm move toward the best solutions.

why we need to cluster tweets?

Most social websites like Facebook and Twitter permit users to post short message on their front page . The messages updated are called status update and this process is termed as micro-blogging [2] .Messages posted on twitter are referred as tweets. Tweets may include videos ,photos and 140 character of text and link etc . Tweets may associated with events like dance ,music , personal views and thinkings .Recently tweets achieved lot of significance because of their capability to spread information rapidly. Most of the famous search engines add these twitter messages in their search results because large no of tweets are adding on each day and contain useful information. So analyzing micro-blogging system is the area where most of the researcher works.

- Tweets are of 140 characters and hence very small in length . This short messages gives very little information about the specific subject .
- Tweets includes short forms, net-slangs,abbreviations,grammer mistakes ,fragments of sentences etc .So it is complex to extract the information from tweets because of their informal style

Clustering of twiiter dataset help to categorize the tweets based upon the content . By using clusters it is easy to identify the particular event or topic about which tweet is written .To compute the similarity of text words for any specific event data is represented by vector space model using term frequency and inverse term frequency. So Particle swarm optimization (PSO) technique is the best method to solve clustering problem.

METHODOLOGY

1. Upload the Twitter data sample
2. Apply PSO algorithm to get the clusters.
   At the initial stage [1], each particle randomly chooses k numbers of document vectors from the document collection as the cluster centroid vectors.

For each particle: (a) Assigning each document vector in the document set to the closest centroid vector.
(b) Calculating the fitness value .
(c) Update velocity and particle position to genrate the next solutions.

Repeating step until one of following termination conditions is satisfied.

(a) The maximum number of iterations is exceeded or

(b) The average change in centroid vectors is less than a predefined value

3. Optimize the parameters like: Clustering size, No of iterations, Passes, Velocity Of particle, Position of particle, No of particles.

All of the above steps are performed in JAVA environment.

STEP 1- Initialize all particle.
STEP 2- for each particle do
STEP 3- Calculate fitness value of particles
STEP 4- If fitness value better than previous pbest then
STEP 5- Set fitness value as new pbest
STEP 6- Choose the particle with best pbest as gbest
STEP 7- Calculate particles velocity
STEP 8- Update particles position
STEP 9- Repeat
STEP 10-Go to step 2
STEP 11-Untill stop by user or maximum iteration is not achieved
STEP 10- Generate Graphs

PSO optimizes problem with large no of candidate solutions . Major objective of this research is to investigate possibilities for the improvement of the effectiveness of document clustering, by finding out the main reason of ineffectiveness of the already built algorithm and to get their solution .The proposed technique will be implemented on seven documents used for clustering

PARAMETERS OF EVALUATION

Cluster size swarm size, number of iterations, velocity components etc are some of the PSO algorithm parameter .the performance of algorithm is affected by these parameters .some of the parameters have great effect on the performance or efficiency of PSO while some of them have little impact.

Cluster Size

Define the size of each cluster .In this research we set cluster size 3 and minimum cluster size should be 2.

Particles Size

Number of particles *n* in the swarm define the size of Swarm. larger no of solution(particles), maximum search space to be covered .in addition if we use large amount of particles then it is more time consuming and computational complexity will also increase. From the research it has been demonstrated that swarm size lies in the range of N [20-60] is used in PSO implementations .In this research we use 4 particles**.**

No of iterations or passes

No of iterations means passes. How many passes are applied to solve a problem to get the effective solution depend upon the problem .Search process may be stopped if less iteration applied ,in second case large no of iteration increase the complexity and is time consuming.

Passes have two parameters on which it works and they are:-

- Velocity of Particle
- Position of Particle

Velocity Of particle

The count at which particle move in problem space. Velocity of particle depends upon the velocity component these velocity components are very important for updating particle's velocity.Velocity component needed to update the velocity of particle.

Three velocity components are used for updating the velocity they are:-

- Inertia component
- Cognitive component
- And Social Component

- **Inertia component**

This component remember the past movement of particle i.e. previous movement of particles through which it moves to next .Inertia component represent immediate change toward next position. The term $\mathbf{v_{ij}}^{\mathbf{t}}$ is termed as inertia component.

Cognitive component

This component compares and measures the particles performance with its past performance.

$\mathbf{c_1 r_1}^{\mathbf{t}}\ \mathbf{[P^t_{best,i} - x^t_{i,j}\ ]}$ is called cognitive component.

Social Component

This component is used to compute particles performance with its neighbors'.

$\mathbf{c_2 r_2}^{\mathbf{t}}\ \mathbf{[G_{best} - x^t_{i,j}]}$ is called social component

Position of particle

The count at which particle move in problem space. In each iteration every time velocity and position updated .After the updation of velocity and position again fitness value of particle is calculated.

- pfbest - particle personal best fitness value
- *gfbest - particle global best fitness value
- pbest - particle personal best fitness value's position

Acceleration coefficients

Properties of acceleration coefficient are [3][5]

c1=c2=0	When these two acceleration coefficient are equal then all particle travels with their current speed so from velocity update equation $\mathbf{v_{ij}}^{\mathbf{t+1}} = \mathbf{v_{ij}}^{\mathbf{t}}$
c1>0 c2=0	Particle in whole swarm are independent and the velocity equation will be $\mathbf{v_{ij}}^{\mathbf{t+1}} = \mathbf{v_{ij}}^{\mathbf{t}} + \mathbf{c_1 r_1}^{\mathbf{t}}\ \mathbf{[P^t_{best,i} - x^t_{i,j}\ ]}$
c2>0 c1=0	This property describe that every particle in swarm is attracted toward single point and velocity equation will be $\mathbf{v_{ij}}^{\mathbf{t+1}} = \mathbf{v_{ij}}^{\mathbf{t}} + \mathbf{c_2 r_2}^{\mathbf{t}}\ \mathbf{[G_{best} - x^t_{i,j}]}$
c1=c2	all particles are attracted towards the average of $\mathbf{P^t_{best,i}}$ and $\mathbf{G_{best}}$
c1>>c2	Each particle is more strongly influenced by its personal best position, resulting in excessive wandering.
c2>>c1	Then all particles are much more influenced by the global best position, which causes all particles to run prematurely to the optima [

Table 1.1: Shows the properties of acceleration coefficient

The effect of velocity components is managed by acceleration coefficients and random values. Here

Acceleration Coefficients **c1** represent the confidence of particle with itself.
Acceleration Coefficients **c2** represent the confidence of particle with its neighbors.

Equation to calculate velocity of each particle[3]

$$v_{ij}^{t+1} = v_{ij}^{t} + c_1 r_1^{t} [P^{t}_{best,i} - x^{t}_{i,j}] + c_2 r_2^{t} [G_{best} - x^{t}_{i,j}] \quad \text{..... (eq. 1.1)}$$

For gbest PSO method, the velocity of particle is calculated by using this equation
Where
t Here term 't' represent time.
i Term 'i' stands for particle .
j Term 'j' is used to denote Dimension.
v_{ij}^{t} Velocity vector of particle 'i' in dimension 'j' at time't'
$x^{t}_{i,j}$ Position vector of particle 'i' in dimension 'j' at time t
$P^{t}_{best,i}$ Personal best positions of particle 'i in dimension 'j'.
G_{best} It is the particle 'i's global best positions in dimension 'j'.
c_1 *and* c_2 are the positive acceleration constant.
r_1^{t} *and* r_2^{t} These are the random numbers used by pso which is(0,1)
$c_1 r_1^{t} [P^{t}_{best,i} - x^{t}_{i,j}]$ it is the cognitive component of velocity.
$c_2 r_2^{t} [G_{best} - x^{t}_{i,j}]$ It is the social component of velocity.

Equation to find the weight vector of each document:
Weight vector of each document is calculated by using this equation (1.2)

$$w_{ji} = tf_{ji} \times idf_{ji} = tf_{ji} \times \log_2 (n / df_{ji}) \text{.... (eq. 1.2)}$$

Where
tf_{ji}- Represent frequency of term i.e. it count the existence/occurrence of term in document. Where 'i' represent the term and 'j' represent the document.
idf_{ji}- Indicates frequency of term in all documents.
n- Represent whole/total document.
Example if there are 7 document in database and document 1 contain text (engineering is the application of scientific)if we want to calculate the word engineering's weight then we have to first find the occurrence of word in all documents for example engineering exist in 3 times in all documents .Then weight vector for engineering is calculated by equation (1.2)

$$\frac{\textit{log of total no of documents}}{\textit{occurance of that word in diffrent document}} \quad \text{... (eq. 1.3)}$$

Example

$$\log_2 \frac{7}{3} = \log \frac{\frac{7}{3}}{2} = \frac{0.36797699}{0.30103} = 1.222392$$

Equation to find the fitness value
Fitness value depend upon cosine similarity which can be calculated using equation 1.4

$$\text{Cosim}(\overrightarrow{t1}.\overrightarrow{t2}) = \frac{\overrightarrow{t1}.\overrightarrow{t2}}{\|\overrightarrow{t1}\|\|\overrightarrow{t2}\|} \quad \text{..... (eq.1.4)}$$

example
t1 (1, 1, 0, 1) and t2 (2, 0, 1, 1)

$$\text{Cosim} = \frac{1*2+1*0+0*1+1*1}{\sqrt{1^2+1^2+0^2+1^2}*\sqrt{2^2+0^2+1^2 1^2}} = 0.72$$

After calculating the cosin similarity value the PSO algorithm compare the best cosin
similarity value and give the best particle which is nearest to the solution. Untill termination conditions are satisfied this steps repeated to get the best accuracy.

Geometrical Representation of PSO[3]

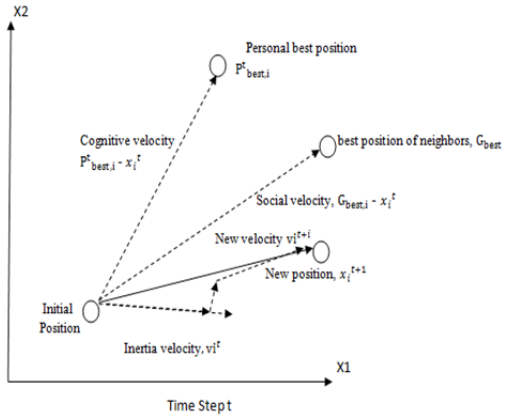

Figure1.1: Velocity and position update for a particle in 2-D search space at time t.

Figure demonstrate how three velocity components helps to move the particle towards the global best position at time steps t respectively

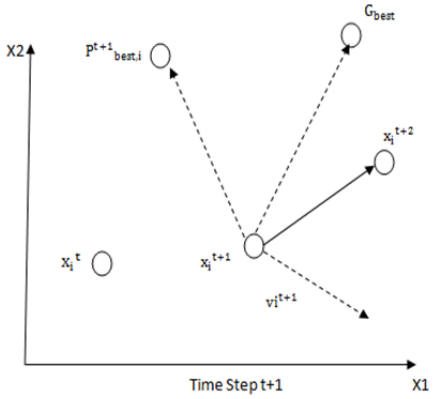

Figure1.2: Velocity and position update for a particle in a two-dimensional search space at time t+1.

This figure demonstrate how three velocity components helps to move the particle towards the global best position at time step t+1 respectively

RESULTS AND DISCUSSION

The whole simulation is taken place in JAVA environment and the following result snapshot shows that proposed method has been tested and following results has been found in which f-score, clustering accuracy, precision and number of iterations taken has been shown.The following formula is used for calculating the accuracy of the desired system[6].

$$\textit{Accuracy} = \frac{\textit{Total correctly classified data}}{\textit{Total sample to classify}}$$

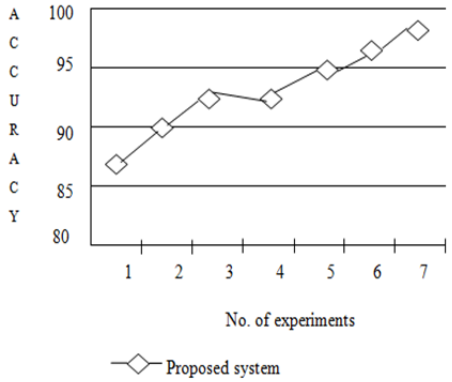

Figure 1.3 Clustering Performance

CONCLUSION AND FUTURE SCOPE

Research paper present document clustering technique using PSO. In this, the global searching area and local refining area are two areas in which behavior of clustering can be described. Here global searching means broad area that all particles tries to cover to find best solution.If the dataset is large then PSO uses global search strategy to get the optimum result of clustering. This global searching ability of the PSO algorithm is one of important technique in PSO and it avoids K means algorithm's limitations.Our experimental result show that higher compact clustering can be generated using this PSO algorithm than using the K-means alone .Many real life problem domains are where Particle Swarm Optimization (PSO) is used . When compared with other algorithms this is easy to implement and needs only less parameter. This can be utilized to solve optimization problems.

There is an extraordinary and good execution performance which is one of the important application approch evolving in document clustering. In complex problems PSO gives better outcome because in PSO there are only few parameters which can be easily adjusted and gives bugs free results with high speed. This feature of PSO makes it widely used technique for optimization.

In future we can concentrate on these:-

- In future better result can be acquired by merging PSO with another methods of optimization.
- To optimise non convex problems we can use this.
- Topology selection and proper parameter are few terms in which further research can be done.For further research there is good scope.

REFERENCES

[1] Xiaohui Cui, Thomas E. Potok ,'' **Document Clustering using Particle Swarm Optimization**'' *Applied Software Engineering Research Group Computational Sciences and Engin-eering Division Oak Ridge National Laboratory Oak Ridge, TN 37831-6085*

[2] Karandhikar Anand ,**clustering short status messages :A topic model based approach**" *university of Maryland ,Baltimore Country,(July 2010)*

[3] Satyobroto Talukder, **" Mathematical Modelling and Applications of Particle Swarm *Optimization*''***School of Engineering Blekinge Institute of Technology SE – 371 79 Karlskrona Sweden,2010***

[4] Andries P. Engelbrecht, **"Computational Intelligence: An Introduction"**: *John Wiley and Sons, 2007, ch. 16, pp. 289-358.*

[5] F. van den bergh, **" An Analysis of Particle Swarm Optimizers",** *Department of Computer Science., 2006, University of Pretoria, South Africa.*

[6] Ruchika Mavis Daniel, Arun Kumar Shukla, **" Improving Text Search Process using Text Document Clustering Approach'',** *"International Journal of Science and Research (IJSR)" 2012*

[7] Jayshree Ghorpade-Aher, Vishakha Arun Metre, **" PSO based Multi-dimensional Data Clustering: A Survey**" *International Journal of Computer Applications (0975 – 8887) Volume 87 – No.16, February 2014*

[8] Jain, A.K, Murty, M.N., and Flynn P.J. 1999. **" Data clustering: a review.** *ACM Computing Surveys, pp. 31, 3, 264-323.*

[9] J. Han and M. Kimber. 2000. **"Data Mining: Concepts and Techniques**. *Morgan Kaufmann.*

[10] Jain, A.K, Murty, M.N., and Flynn P.J. 1999. **"Data clustering: a review. ACM Computing Surveys"**, pp. 31, 3, 264-323.

[11] M. Steinbach, G. Karypis, and V. Kumar. 2000. **" A comparison of document clustering techniques".** *KDD Workshop on Text Mining.*

[12] P. Berkhin. 2004. **"Survey of clustering data mining techniques** [Online].Available:http://www.accrue.com/products/rp_cluster_review.pdf.

[13] XuRui. 2005. **"Survey of Clustering Algorithms**. *IEEE Transactions on Neural Networks, 16(3):pp. 634-678.*

[14] L. Zhuang, and H. Dai. 2004**. "A Maximal Frequent Itemset Approach for Document Clustering.** *Computer and Information Technology, CIT. The Fourth InternationalConference, pp. 970 – 977.*

[15] R. C. Dubes and A. K. Jain. 1998**. " Algorithms for Clustering Data".** *Prentice Hall collegeDiv, Englewood Cliffs, NJ, March.*

[16] D. Koller and M. Sahami. 1997. **" Hierarchically classifying documents using very few words"**. *In Proceedings of (ICML) 97, 14th International Con-ference on Machine Learning, pp. 170–178, Nashville, US.*