International Conference On Emanations in Modern Technology and Engineering (ICEMTE-2017) ISSN: 2321-8169
Volume: 5 Issue: 3 176 - 178

Semantic Subgraph Generation for Hindi Text Document using PSO

Ragini Mishra
Department of Computer Engineering
University of Mumbai
SLRTCE, Mumbai , India
raginimishra567@gmail.com

Abstract **-** In present scenario, numerous technology and thesis are developing and lots of analytical researches are in progress with the help of current and previous experimental documentation. This resulted in huge amount of data that needs to be minimized so as to discovery the information rooted in such documents. This paper presents, the semantic subgraph generation for Hindi text document using Particle Swarm Optimization (PSO. The goal is to extract the important sentences for the summary from the Hindi text document. PSO algorithm is used for clustering of sentences.

Index Terms – *Text summarization, preprocessed data, clustering, Particle Swarm optimization*

*****

I. INTRODUCTION

Text summarization refers to process of generating summary in precise for the given text document. On the basis of syntax and semantics the summary generated by human and computer varies in terms of understanding the depth of document. The process of automatic summarization requires machine to understand the language in depth. The text summarization presented in this paper is based on Extraction method.

A. Text Summarization Techniques

Text summarization refers to process of generating summary in precise for the given text document. On the basis of syntax and semantics the summary generated by human and computer varies in terms of understanding the depth of document. The process of automatic summarization requires machine to understand the language in depth. The text summarization can be classified on the basis of summarization purpose as Indicative or informative or critical summary, on the basis of summarization form as Extractive or Abstractive summarization, on the basis of dimension as Single document summarization and Multi-document summarization and on the basis of context as Query specific or Query independent.

Indicative summary provides a description of what paper covers without hunting into its substance, whereas informative summary provides complete information with reference to subject that paper covers.

Extractive summarization refers to extract key textual elements like keywords, paragraphs, clauses, sentences or phrases from text using linguistic and statistical analyses. Extractive summarization technique identifies the most relevant sentences from the text document and may be used as summary. Abstractive summarization technique uses linguistic method to understand and describes the text document in few words. In abstractive summarization the summary generated may or may not contain sentences from the original text document but it conveys the most important information from the text document. It provides the concise information about the subject matter in the presented text document.

Single document summarization refers to process of extracting information from one single document. Multi-document summarization refers to process of extracting information from multiple documents on same subject domain.

Query specific summary refers to process of producing summary from the document in reference to user query whereas Query independent summary attempts to identify information in text without the context of a query.

B. Particle Swarm Optimization

Swarm Intelligence is innovative artificial intelligence based on collective behaviour of decentralized self-organized system. It helps in solving optimization problem that originally is inspired from biological examples of swarm theory.

Particle Swarm Optimization (PSO) integrates the behaviour observed in bird flocking or fish schooling and swarms of bees. PSO is made up of the population of simple agents interacting with nearby agent, and with their environment. PSO is evolutionary algorithm used as global optimization tool, which can be used to solve problem in an n-dimensional space. Doctor Kennedy and Eberhart in 1995 [1] presented a heuristic global optimization method called the particle swarm optimization. The birds are either scattered or go collectively in search of food to trace the place where they can find the food. The birds can smell the food very well as they have better food resource information. The birds transmit the information while searching for the food to one another any time. Eventually, the birds flock to the place where the food can be found. The particle swarm optimization algorithm is compared to the bird swarm that moves from one place to another in search of food resource is equivalent as to reach most optimal solution during the course.

PSO [2] starts with random initialization of a population (swarm) of individuals (particles) in the n-dimensional search space (n is the dimension of problem in hand). The particles fly over search space with adjusted velocities.

In PSO, each particle keep two values in its memory:

(1) Its own best experience, that is, the one with the best fitness value (best fitness value corresponds to least objective value since fitness function is conversely proportional to objective function), whose position and objective value are called P_i and P_{best}, respectively
(2) The best experience of the whole swarm, whose position and objective value are called P_g and g_{best}, respectively.
Let denote the position and velocity of particle I with the following vectors:

$X_i = (X_{i1}, X_{i2},\ldots, X_{id},\ldots ,X_{in})$

$V_i =(V_{i1}, V_{i2},\ldots, V_{id},\ldots ,V_{in})$

The following equation for PSO updates the velocity and position of the particle [2]:

$$V_{id}(t+1)= V_{id}(t) + c_1*r_{1d}*(P_{id} - X_{id}) + c_2*r_{2d}*(P_{gd} - X_{id}) \quad (1)$$

$$X_{id}(t+1)= X_{id}(t) + V_{id}(t+1)$$

Where,
$V_{id}(t)$ is velocity of particle i at iteration t.
$X_{id}(t)$ is position of particle i at iteration t.
$V_i(t+1)$ is velocity of particle i at iteration t+1.
Xi (t+1) is velocity of particle i at iteration t+1.
r1d, r2d is random number between (0,1).
c1 cognitive acceleration coefficient.
c2 social acceleration coefficient

II. PSO Clustering Algorithm

Proposed System

1. The algorithm accepts the preprocessed data in order to generate the summary for the given input data (Choose Dataset).
2. The algorithm initializes particle with random velocity, position and inertia then calculates the fitness value for each of the particle (Start Process).
3. The algorithm iterates till it obtains the pbest value better than the previous one for each of the particles. On receiving the better value it replaces the pbest with the new value else pbest value remains the same and updates it as a gbest of each of the particles respectively.
4. The velocity of the individual particles are calculated and updated to it data value.
5. The PSO algorithm converges only when there is no change in value of gbest of particle for maximum 10 iterations. If there is no change the process is repeated till it meets stopping criteria.
6. Thus, the obtained particle is included to be a part of summary.

How proposed system clusters the summary sentences

1. The algorithm distributes all the data points to K cluster for each of the particle and iterates over all data points to distribute them to the respective cluster.
2. Euclidean distance between all the centroid of the particle is evaluated and assigned a cluster index (i.e. 1 or 0).
3. After all iterations intra and inter cluster distance is obtained based on which the final centroid of gbest particle is obtained.
4. This distance is calculated using Euclidean distance.

III. Implementation

The GUI comprise of an option for selecting Dataset i.e. Sample Dataset and Training Dataset. Once the Dataset is chosen click on Start Process button in order to begin with algorithm execution. The button called Clear Screen used to flush the content on the output screen. The result is displayed on the Output of Algorithm window that shows the final gbest centroid of the given dataset. The sample data set is used to test the PSO clustering algorithm for small dataset as shown in the fig 5.1 where the dataset of length 490 is used to test the PSO Clustering algorithm

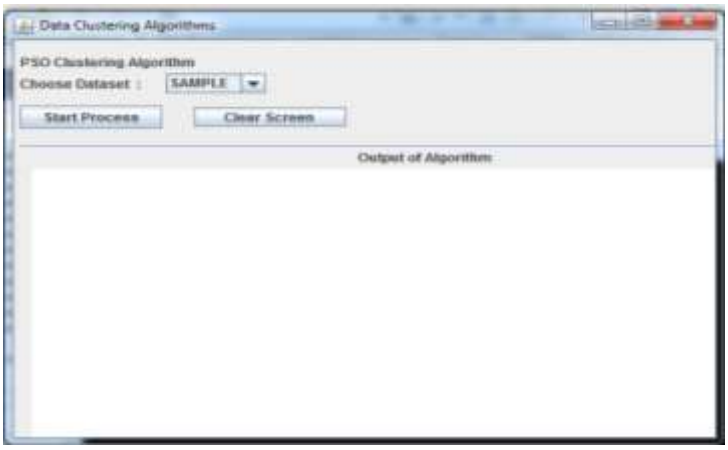

Fig. 1 User Interface for PSO Clustering Algorithm of Sample Dataset

The output is displayed within few seconds for the small dataset. Thus the performance is good and algorithm converges in few iterations.

Fig 2. Centroid based Summary of Sample Dataset

The Training dataset selected consist of pre-processed 30 new articles when given to the PSO clustering algorithm takes approximately three minutes to complete the process and display the results. This is due to the length of dataset which is 946 and takes more time to converge.

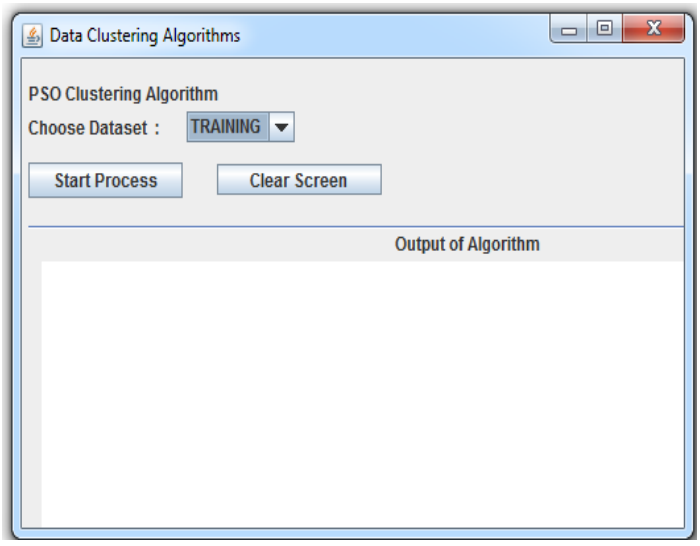

Fig 3. User Interface for PSO Clustering Algorithm of Training dataset

The output for Training dataset takes more time to converge due to large dataset size. Thus generates the final centroid of the gbest particle.

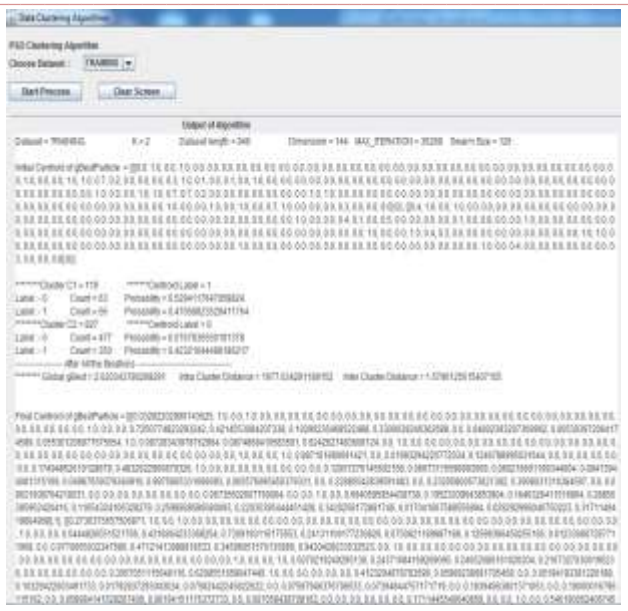

Fig 4. Centroid based Summary of Training Dataset

CONCLUSION

Particle swarm optimization (PSO) algorithm that is an effective technique and its performance depends on the PSO variants and parameter selection. In order to achieve 100% accuracy using PSO as a classifier is uncertain and the convergence time is also unpredictable. So we have analysed the classifier using problem dimension of 100 and to achieve the stopping criteria pre-specified number of iteration (i.e. 10 iterations) is considered. To further refine the input document, we can integrate the co-reference resolution in preprocessing stage.

REFERENCES

[1] J. Kennedy and R. Eberhart, “Particle swarm optimization,” Proceedings IEEE International Conference Neural Networks, 1995, pp. 1942–1948
[2] Rezaee Jordehi, A., and J. Jasni. "Parameter selection in particle swarm optimisation: a survey." Journal of Experimental & Theoretical Artificial Intelligence 25.4 (2013): 527-542.