International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Research Article

Text Clustering using PBO algorithm for Analysis and Optimization

Manpreet Kaur[Á*] and Navpreet Kaur[Á]

[Á]Assistant Professor , Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India

Accepted 05 Nov 2014, Available online 01 Dec 2014, **Vol.4, No.6 (Dec 2014)**

Abstract

Text clustering refers to divide text collection into small clusters and require similarity as large as possible in same cluster. Textual clustering technique was introduced in the area of text mining. The two important goals in text clustering are achieving high performance or efficiency and obtaining highly accurate data clusters that are closed to their natural classes or textual document cluster quality. In order to obtain useful information quickly and accurately form the mass information, text clustering technique is an important research direction. The k-means clustering algorithm has limitations, which depends on the initial clustering center and needs to fix the number of clusters in advance. For these reasons a text clustering algorithm based on latest semantic analysis and optimization is proposed. Thus, a new clustering algorithm based on PBO and optimization has been proposed, which effectively solved the high dimensional and sparse problem and overcomes the dependency of the number of clusters and initial clustering center of k –means algorithm.

Keywords: *Text clustering, Latent Semantic Analysis, K-means clustering algorithm, clustering optimization, GA, PSO*

1. Introduction

Data mining is defined as the procedure of handling data from different sources and summarizing it into useful information. Knowledge Discovery in Data is another name for data mining (Bhavani Thuraisingham *et al*, 2012). Text mining involves the process in which the input text is structured (usually parsing, involving the addition of some derived linguistic features and the removal of others, and their subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. Mostly text mining tasks include text categorization, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling, Concept /entity extraction, text clustering.

2. Clustering in Text Mining

In recent years, many different optimization techniques have been proposed for solving the complex, multimodal functions in several fields. Some of them well-known optimization algorithms are the genetic algorithm (GA), Particle swarm optimization (PSO) algorithm. These algorithms are used by many researchers to obtain the optimum value of the problems.

The objective of clustering is to partition an unstructured set of objects into clusters (groups). One often wants the objects to be as similar to objects in the same cluster and as dissimilar to objects from other clusters as possible.

*Corresponding author **Manpreet Kaur** and **Navpreet Kaur** are M.Tech, Research Student

To use most clustering algorithms two things are needed:

- An object representation,
- A similarity (or distance) measure between objects.

Clustering and classification are both fundamental tasks in Data Mining (Canyu Wang, Xuebi Guo, Hao Han *et al*, 2012;Wang Chun-hong , Nan Li-Li; Ren Yao-Peng *et al*, 2011). Classification is used mostly as a supervised learning method, clustering for unsupervised learning (some clustering models are for both). The goal of clustering is descriptive, that of classification is predictive (Canyu Wang, Xuebi Guo, Hao Han *et al*, 2012). Since the goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is intrinsic. Clustering group data instances into subsets in such a manner that the instances which are similar grouped together, while different instances belong to different groups (Wang Chun-hong , Nan Li-Li; Ren Yao-Peng *et al*, 2011). The instances are thereby organized into an efficient representation that characterizes the population being sampled (Canyu Wang, Xuebi Guo, Hao Han *et al*, 2012;Wang Chun-hong , Nan Li-Li; Ren Yao-Peng *et al*, 2011).

3. Enhancements on Clustering Algorithm

The text clustering based on Vector Space Model has problems, such as high dimensional and sparse, unable to solve synonym and polyseme etc. And meanwhile, k-means clustering algorithm has shortcomings, which depends on the initial clustering center and needs to fix the number of clusters in advance. For this, a text clustering

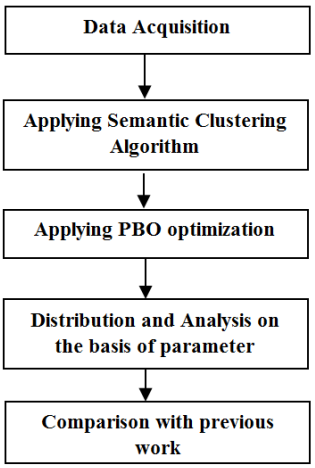
algorithm based on Latent Semantic Analysis and optimization is proposed. This algorithm can not only overcome the problems of Vector Space Model, but also can avoid the shortcomings of k-means algorithm. As compared with the text clustering algorithm based on LSA and text clustering algorithm based on VSM and optimization (Wang Chun-hong , Nan Li-Li; Ren Yao-Peng *et al*, 2011).

A hybrid Particle Swarm Optimization (PSO) + K-means document clustering algorithm that performs fast document clustering and can avoid being trapped in a local optimal solution as well. For comparison purpose, we applied the PSO+K-means, PSO, K-means, and other two hybrid clustering algorithms on four different text document datasets. The number of documents in the datasets range from 204 to over 800, and the number of terms range from over 5000 to over 7000. The results illustrate that the PSO+K-means algorithm can generate the most compact clustering results than other four algorithms (Xiaohui Cui, Thomas E. Potok *et al*, 2005).

4. Proposed Work

Pollination Based Optimization

Optimization is defined as a natural process embedded in the biotic beings. Pollination is a process involving the transfer of pollen from male parts of flower named anther to the female part named stigma of a flower. Often some majority of flowers develops seeds as a result of self-pollination, when pollen and pistil are from the same plant, often but not always from the same flower. While other plants need cross-pollination pollen and pistil are essentially from different plants. However, pollinators have no concern about plants benefits. They just pollinate in search of nectar and/or pollen from flowers to meet their energy requirements and to produce offspring. In this research clustering is done for optimization of K-mean clustering algorithm using PBO algorithm.



Basic Design of Proposed Work

The detailed procedure of this algorithm is as following

Step 1: Initialize PBO Parameters.
Plants – Number of clusters
Seasons –Iterations
Weeks – Number of data it include
Step 2: Evaluate Reproduction factor according to clusters and static parameters change the fitness or R according to iteration or season. Text clusters is changing according to value initialized.
Step 3: Clusters optimized in accordance with R-factor.
Step 4: Start single iterations with every value.
Step 5: α, β, γ value as assumptions.
Step 6: Check efficiency with every dataset and including clusters.
Step 7: Main value of clusters must not repeat.
Step 8: End.

5. Experiment Result and Analysis

To prove the effectiveness of the text clustering algorithm based on PBO and optimization, we compare it with GA.

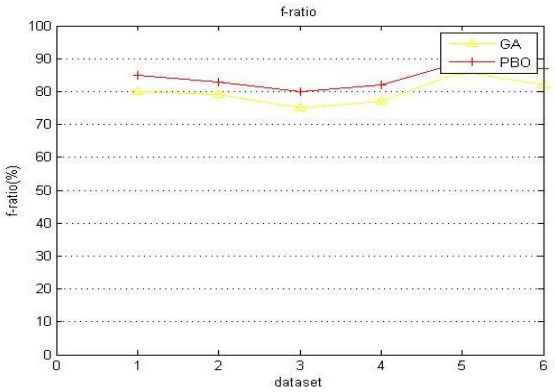


Fig 1: Representation of graphs for f-ratio.

In fig. 1 it describes that dataset is taken to the f-measure. Here, red line is for PBO and yellow is for GA. In this graph PBO is better because higher the value of f-ratio higher the efficiency.

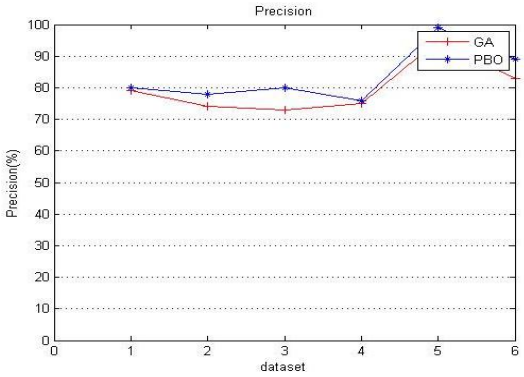


Fig 2: Representation of graphs for Precision.

In fig. 2 it depicts that dataset is taken to the f-measure. Here, blue line is for PBO and red is for GA. In this graph PBO is better because higher the value of precision higher the efficiency.

In fig. 3 it represents that dataset is taken to the f-measure. Here, red line is for PBO and blue is for GA. In this graph PBO is better because higher the value of recall higher the efficiency.

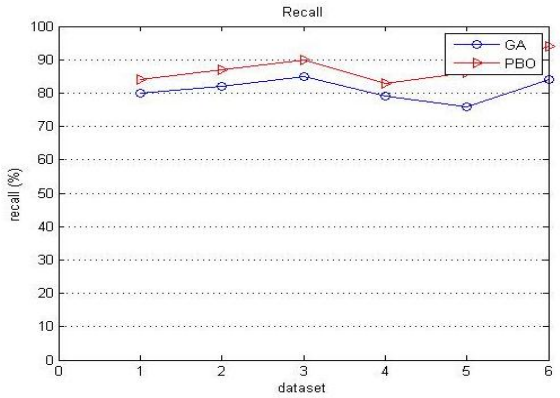


Fig 3: Representation of graphs for recall.

Table 1 List of Parameters

Iterations	Parameters					
	Precision		Recall		F-Measure	
	GA	PBO	GA	PBO	GA	PBO
1	79	80	80	84	80	85
2	75	79	82	88	79	82
3	72	80	86	90	75	80
4	76	76	79	82	77	82
5	92	98	76	84	86	92
6	83	90	85	95	82	88

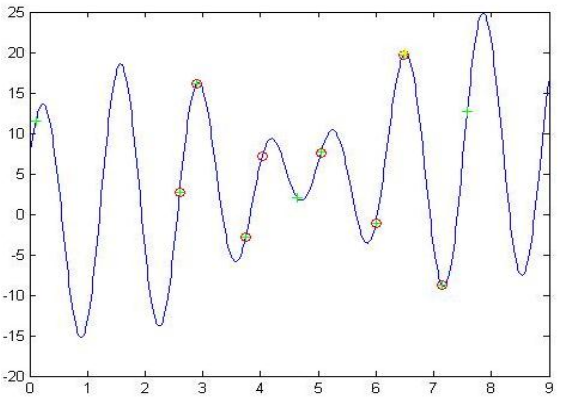


Fig 4: Representation of optimized clusters

In fig. 4 the circles here are optimized clusters the graph is generated for text clustering.

6. Performance Analysis

Precision: In simple language precision is the fraction of retrieved documents that are relevant to the search, for a text search on a set of documents precision is the number of correct results divided by the number of all returned results.

Precision= (Number of relevant documents retrieved)/ Total number of documents retrieved.

Recall: Recall is the fraction of the documents that are relevant to the query that are successfully retrieved, for text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned.

Recall = (Number of relevant documents retrieved)/ (Total number of relevant documents)

F-RATIO: A measure that combines precision and recall is the harmonic mean of precision and recall.

F-ratio = 2 * ((Precision. Recall) / Precision + Recall).

Conclusion

In this research clustering is done for optimization of K-mean clustering algorithm using PBO algorithm. Now we can analyze for semantic clustering PBO algorithm. It gives better result than the previous algorithm and this can be concluded on the bases of some parameter like recall, precision and f-measure. This can give better results for the conceptual clustering. And experiments prove that our algorithm greatly improves the effect of text clustering, and then provides a more forceful and better support for text efficient retrieval.

Future Work

In future, this can be enhancing by implementing some hybrid algorithm. The other potential areas of future work include comparison of the proposed scheme with other artificial intelligence techniques or another artificial intelligence technique is used for optimization.

References

M. Praveen, Dora Babu Sudarsa (2013), A Customized Vector Space Model Implementation in Document Clustering to Enhance the Performance International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 5.

Canyu Wang, Xuebi Guo, Hao Han (2012), Crime Detection Using Latent Semantic Analysis and Hierarchical Structure IEEE 3rd International Conference on Software Engineering and Service Science (ICSESS), pp. 337 – 340.

Wang Chun-hong , Nan Li-Li; Ren Yao-Peng (2011), Research on the text clustering algorithm based on latent semantic analysis and optimization IEEE International Conference ,vol.4, pp. 470 – 473

Xiaohui Cui, Thomas E. Potok (2005), Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm Journal of Computer Sciences, pp. 27-33.

Bhavani Thuraisingham (2012), Data Mining for Malicious Code Detection and Security Applications-IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology – Workshops .