AUTOMATED INFORMATION RETRIEVAL MODEL USING FP GROWTH BASED FUZZY PARTICLE SWARM OPTIMIZATION

Raja Varma Pamba[1] Elizabeth Sherly[2] and Kiran Mohan[3]

[1]School of Computer Sciences, Mahatma Gandhi University, Kottayam,India
[2]Indian Institute of Information Technology and Management-Kerala, Trivandrum,India
[3]Payszone LLC LTD, Dubai,UAE

ABSTRACT

To mine out relevant facts at the time of need from web has been a tenuous task. Research on diverse fields are fine tuning methodologies toward these goals that extracts the best of information relevant to the users search query. In the proposed methodology discussed in this paper find ways to ease the search complexity tackling the severe issues hindering the performance of traditional approaches in use. The proposed methodology find effective means to find all possible semantic relatable frequent sets with FP Growth algorithm. The outcome of which is the further source of fuel for Bio inspired Fuzzy PSO to find the optimal attractive points for the web documents to get clustered meeting the requirement of the search query without losing the relevance. On the whole the proposed system optimizes the objective function of minimizing the intra cluster differences and maximizes the inter cluster distances along with retention of all possible relationships with the search context intact. The major contribution being the system finds all possible combinations matching the user search transaction and thereby making the system more meaningful. These relatable sets form the set of particles for Fuzzy Clustering as well as PSO and thus being unbiased and maintains a innate behaviour for any number of new additions to follow the herd behaviour's evaluations reveals the proposed methodology fares well as an optimized and effective enhancements over the conventional approaches.

KEYWORDS

Information retrieval, Clustering, Fuzzy particle swarm optimization & Frequent Pattern growth

1. INTRODUCTION

Huge influx of information in the web poses the greatest challenge before all researches to find an effective way to find the best of all relevant information matching the search context. Searching and indexing methods in retrieval systems exists in multitudes but one that finds information matching to the relevance to the search query automatically still remains a problem unresolved. Web documents exhibit the property of being belonging to more than one cluster. This quality fine tunes to algorithms that caters to partitional clustering out performing hierarchical clustering. Hierarchical clustering fails at situations when it needs to trace back to new clusters if any changes happens. While in partitional clustering adapts well to soft clustering. The traditional approaches K means and FCM much in popular fails with regard to initial supply of random cluster centroids by making the system biased to the user inputs. In the proposed methodology to my knowledge first of its kind to make the system by itself generating meaningful inputs to fuel the Fuzzy Particle swarm optimization. Except for the initial set of documents to be clustered no other inputs are given to the system making the system automated.

DOI:10.5121/ijcsit.2017.9109

The rest of the paper is organized as follows. Section 2 deals with the related works. Section 3 describes the problem and the research objectives. Section 4 discusses Proposed Methodology. The Experimental results and Conclusions are described by Section 5 and 6 respectively.

2. RELATED WORK

Paper by Liu.et.al (2008) discusses on the combination of PSO and Fuzzy C-Means . There is also an integration of PSO and K-means algorithm (KPSO). Yau.et.al (2013)improves KPSO algorithm by proposing an enhanced cluster matching.In the authors proposed an innovative approach of using PSO to overcome the shortcomings of FCM clustering .Statistical analysis based principal component analysis was used a variant of Fuzzy clustering in J. C. Bezdek, C. Coray work.

3. RESEARCH OBJECTIVE

The main objectives of the proposed methodology are find ways to achieve the following objectives

1. Find all possible semantic relatable frequent sets as the reduced search space for the clustering to begin with.
2. To find from among the semantic relatable sets the initial optimal cluster centroids and the set of particles where partitional clustering fails to find it.
3. Retrieve all possible fuzzy clusters matching the user query relevance using Frequent pattern growth based Fuzzy Particle swarm optimization abbreviated as FPFPSO for the rest of the paper.
4. Evaluate the cluster quality for its internal cluster quality.

4.METHODOLGIES

4.1 SEMANTICALLY RELATABLE SETS FOR SEARCH SPACE

4.1.1 FREQUENT PATTERN GROWTH ALGORITHM

In the proposed methodology discussed we make use of Frequent Pattern Growth Algorithm for reducing the search space by retaining all possible semantic relatable frequent item sets matching to the search transactions. With its initial phase dealing with Frequent pattern tree generation we compress all of our keywords into a tree after processioning of documents. In this the keywords are initially checked for the threshold frequency to match with support and confidence to finally prune out all those terms not in relevance to the transaction. By this we retrieve only terms relevant to the search query. Once the FP tree is generated the second stage of FPG is Frequent Pattern Growth for generating all possible semantic relatable frequent items sets. The algorithm retrieves all conditional frequent item sets in use frequently which could for the rest of the paper can be presumed as semantic relatable sets which are related to the search context.

For Example if key terms after pre-processing are Computer, Network, Social then after FP Tree it shows the co-occurrence of each keyword and in every chain of keywords how many times the particular terms have been traversed. With FP Growth, it finally generates all possible combinations like [Computer Network, Social Network, Computer Social Network].} All possible combinations matching the threshold support. A template of the semantic relatable set retrieved from the FP Growth algorithm is as given below in Table 1.

Table 1: Semantic Relatable Frequent Terms set

Keywords[a]	Conditional pattern base	Conditional FP Tree	Generate Frequent [redacted]
A	$<(F,B,E),(F,B,E,D),(C,B,D),(C,F,B,D),(C,F,B,E,D)>$	$<(C:4),(F:5),(B:3),(E:3)>$	$< A, F, A, BA, E,$ [redacted]
D	$<(F,B,E),(C,B),(C,F),(C,F,B),(C,F,B,E)>$	$<(C:4),(F:4),(B:4),(E:2)>$	$< D, C, D, F, D, E$ [redacted]
E	$<(F,B),(C,F,B)>$	$<(C:1),(F:2),(B:3)>$	* [redacted]
B	$<(F),(C),(C,F)>$	$<(C:2,F:2)>$	* [redacted]
F	(C)	$(C:1)$	* [redacted]
C	*	*	* [redacted]

From this those matches the threshold support are filtered and rest are pruned out to reduce the search space. The retrieved bunch has in them all possible semantically related term sets which has higher relevancy to the user search context as shown in Table 1. Indirectly system builds in a concept tagging of highly related terms to form meaning as well as relevance to the search context. By making use of projected data structure based tree generation, a divide and conquer approach strategy to effectively overcome the issue of space constraints of FP Growth algorithm as in [Han J., Pei J., Yin Y. and Mao R.,2003]. The major constraint of every conventional clustering is that the approach is skewed towards the random initialization of inputs. To avoid this the proposed methodology recast the outcomes of FP Growth to suit the demands of FuzzyPSO.

4.2. Fuzzy Clustering

4.2.1. Fuzzy Particle Swarm Optimization

The fuzzy clustering can be considered as an stochastic meta heuristic optimization problem that gives optimized clusters. The objective of optimization is to seek values for a set of parameters that maximize or minimize objective functions subject to certain constraints. Swarm Intelligence is an intelligent paradigm for solving optimization problems that is basically evolved from bio nature inspired mechanisms like fish schooling, bird flocking and swarm behaviour of honey bees. The drawbacks of Fuzzy clustering normally FCM are resolved with the optimization tools like PSO ,ACO and Genetic Algorithms. Out of all evolutionary algorithms the simplest of all in implementation and managing is Particle swarm optimization which does not require any crossover or mutation operators. Web documents being fuzzy and vague the need of the time is call for Fuzzy Particle swarm optimization which caters to the dynamicity and vagueness of web documents.

The fuzzy clustering of objects is described by a fuzzy matrix μ with n rows and c columns in which n is the number of data objects and c is the number of clusters. The element in the i th row and j th column in $\mu_{\{ij\}}$, indicates the degree of association or membership function of the i th object with the j th cluster. The characters of μ are as follows:

$$\mu_{ij} \in [0,1] \forall_i = 1,2...n \forall_j = 1,2...c \tag{1}$$

$$\sum_{j=1}^{c} \mu_{ij} = 1, \forall_i = 1,2,....n \tag{2}$$

$$J_m = \sum_{j=1}^{c} \sum_{i=1}^{n} \mu_{ij}^{m} d_{ij} where d_{ij} = \| o_i - z_j \| \tag{3}$$

$$z_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{m} o_i}{\sum_{i=1}^{n} \mu_{ij}^{m}} \tag{4}$$

$$\mu_{ij} = \frac{1}{\sum_{k-1}^{c}} \frac{d_{ij}}{d_{ik}}^{\frac{2}{m-1}} \tag{5}$$

4.3. HYBRID FP GROWTH BASED FUZZY PARTICLE SWARM OPTIMIZATION

In this proposed methodology as referred to subsection 4.1.1 the parameters for Fuzzy PSO are generated by the system itself after FP Growth algorithm. The set of conditional frequent item sets are considered as the swarm and cluster centroids calculated by finding the average in every sets.

The Fuzzy PSO starts with a population of particles and initial number of clusters obtained from FP growth algorithm whose positions represent the potential solutions for the studied problems and velocities are randomly initialized in the search space. The position matrix X is redefined in this proposed algorithm, represent the fuzzy relation (membership function) between the frequent items sets(particles) in columns and cluster centres as rows. The position matrix is given below:

$$X = \begin{vmatrix} \mu_{11} & \cdots & \mu_{1c} \\ \vdots & \ddots & \vdots \\ \mu_{n1} & \cdots & \mu_{nc} \end{vmatrix}.$$

In each of its iteration, the search for optimal solution is executed by updating the particle velocities and its position. The fitness value of each frequent item sets(particles or swarm) is determined using a fitness function, J_m based on Euclidean Distance measure as defined in Equation 3 .The velocity of each particle is updated using two best position ,individual best position i_{best} and social best solution s_{best} .The individual best position i_{best} is the best position that particle has visited so far and s_{best} is the best position the swarm has visited so far. A particle velocity and position are updated as follows :

$$V(t+1) = w \otimes V(t) \oplus c_1 r_1 \otimes (i_{best}(t) \ominus X(t)) \oplus c_2 r_2 (s_{best}(t) \oplus X(t)) \quad (6)$$

$$K : 1, 2, ..., P \quad (7)$$

$$X(t+1) = X(t) \oplus V(t+1) \quad (8)$$

Where X and Y are position and velocity of particles respectively, w is inertia weight, c_1 and c_2 are constants, called acceleration coefficients which control the influence of i_{best} and s_{best} on the search process, P is the number of particles in the swarm derived from the frequent item sets generated from FP Growth, r_1 and r_2 are random values in the range[0,1].

Based upon the fitness function evaluated, the particle with least fitness function can be eliminated to help improve the search space. Finally the FPFPSO Clustering is executed upon the resultant set to fine tune the elements for clustering.

A stage reaches where the position of the particle saturates with no changes in the position further. Here the algorithm converges to retrieve the final cluster positions of individual documents to its respective clusters as shown in Table 2.

Table 2: Document Cluster Membership position

Document ID's	SET 1	SET 2
1	0.4508929	0.5491071
2	0.4452091	0.5547904
3	0.7772407	0.2227593
4	0.0000000	1.0000000
5	1.000000	0.0000000
6	0.4734385	0.5265615
7	0.0000000	1.0000000
8	0.0000000	1.0000000
9	0.000000	1.0000000
10	0.000000	1.0000000

5. EXPERIMENTAL RESULTS

Conventional approaches KMeans, FCM and FPFCM algorithm have been conducted on Reuters 21578 for the purpose of performance evaluation. In the experiment 50 particles have been trained for 100 iteration. The documents in the Reuters-21578 collection are originally taken from Reuters newswire in 1987. The Reuters-21578 contains 22 files. Each of the first 21 files (reut2-000.sgm through reut2-020.sgm) contains 1000 documents, while the last (reut2-021.sgm) contains 578 documents. The documents are broadly divided into five broad categories (Exchanges, People, Topics, Organizations and Places). These categories are further divided into subcategories. The Reuters-21578 test collection is available at [9]. Figure 1 shows that the fitness value decreases for every iteration and FPFPSO performs better than K-Means, FCM and FPFCM. When the number of clusters increases then the fitness function value is decreased. The K-means algorithm generates a local optima result. Two similar documents in the same cluster are true positive (TP) and two similar documents in different clusters are coined true negative (TN). While two dissimilar documents to the same cluster is false positive (FP) and two similar documents to different clusters is false negative (FN). Precision (P) and Recall(R) are two popular measures for clustering. Thumb rule is higher the precision lower the recall value. The F-measure (F) is used to evaluate the performance measure of the proposed system. A higher F- measure indicates better performance which the proposed methodology achieves as shown in Figure 2. The Precision and Recall are given below. The F-measure is evaluated using the

$$P = \frac{TP}{TP + FP} \qquad (9)$$

$$R = \frac{TP}{TP + FN} \qquad (10)$$

$$F = \frac{2 * P * R}{P + R} \qquad (11)$$

The FPFPSO clustering maximizes the similarity between the documents in the same cluster and also maximizes the distance between the cluster centres. The fitness function of the hybrid algorithm is the objective function of FCM algorithm.

FCM performs well but with random initialization system rather turns biased and hooks at local minima. The bio inspired particle swarm algorithm is a meta stochastic tool that is mostly used to supplement the performance of conventional clustering approaches. The proposed algorithm fares well with the convergence speed of PSO and also overcomes the shortcomings of FCM to escape from local optima.

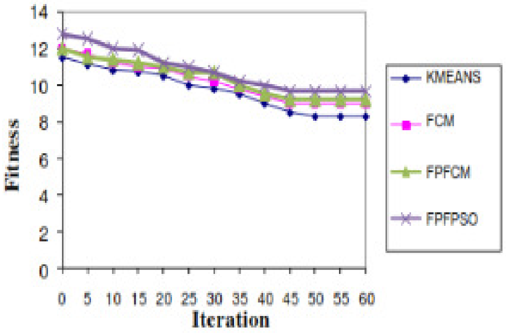

Figure 1: Fitness Value Comparison

Table 3: FMEASURE PERFORMANCE EVALUATION

FMEASURE	KMEANS	FCM	FPFCM	FPFPSO
fmeasure	0..543	0.691	0.867	0.943

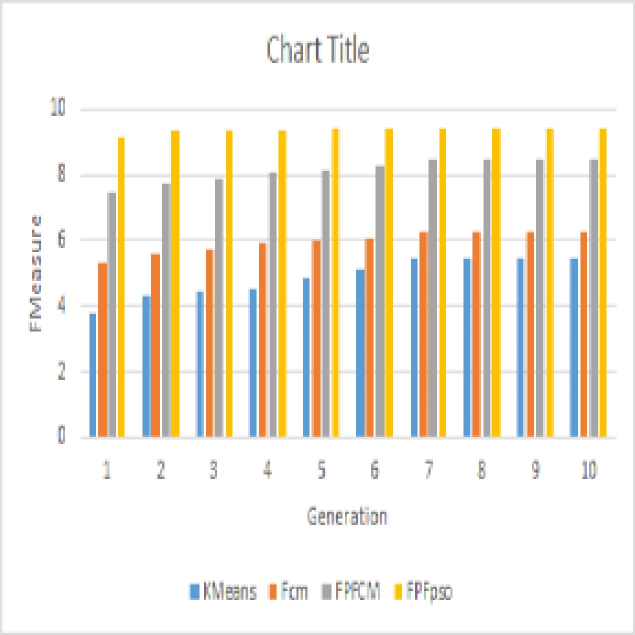

Figure 2: F-Measure Comparison.

6. CONCLUSIONS

In this proposed approach, the methodology takes into account primarily the base issues of any information retrieval systems effectively. To instil concept tagging and relationship in the retrieved patterns, we use FP Growth. Use of FP Growth also helps reduce search space by identifying all transactions relevant to the search context and generates all possible combinations of related terms. From this bunch of highly related cooccurrence data points taken as swarm and optimal number of clusters been taken from FP Growth we direct the same to Fuzzy PSO to inherit and discipline in the system to a specific behaviour, this takes care of attracting the documents to its respective cluster nest by following the Fuzzy PSO. Evaluation results as shown in Table 3 and Fig 1 gives a promising perspective end towards an efficient and effective clustering which minimizes the intra cluster distances and maximizes the inter cluster differences.

REFERENCES

[1] Tan, P.N., Steinbach, M. and Kumar, V. " Cluster Analysis: Basic concepts and algorithms", Introduction to Data Mining, Pearson Addison Wesley, Boston, pp. 487-568, 2006.

[2] Baeza-Yates, R. and Ribeiro-Neto, B. “Modern Information Retrieval”, Addison Wesley, 1999.

[3] Bezdek, J. “Pattern recognition with fuzzy objective function algorithms”, New York, Plenum Press, 1981.

[4] Cui, X., Potok, T.E. and Palathingal, P. “Document clustering using particle swarm optimization”, Proceedings of IEEE Swarm Intelligence Symposium, pp. 185-191, 2005

[5] Abraham, A., Guo, H. and Liu, H. “Swarm intelligence: foundations, perspectives and applications”, Swarm Intelligent Systems, Nedjah, N.and Mourelle,L. Eds. Studies in Computational Intelligence, Springer Verlag Germany, pp. 3-25, 2006.

[6] Han J., Pei J., Yin Y. and Mao R "Mining frequent patterns without candidate generation: A frequent-pattern tree approach",Data Mining and Knowledge Discovery,2003.

AUTHORS

Raja Varma Pamba is an Assistant Professor in the Department of LBS Institute of Technology for Women, Trivandrum, Kerala, India and pursuing Ph.D in School of Computer Sciences, Mahatma Gandhi University, Kerala, India.

Dr Elizabeth Sherly is the Professor & Head for VRCLC department at Indian Institute of Information Technology and Management-Kerala, India .

Kiran Mohan is the Operation Director for Payszone LLC Ltd,Dubai