International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 11 | Nov -2016 www.irjet.net p-ISSN: 2395-0072

Analysis of Document Clustering using Pseudo Dynamic Quantum Clustering Approach

Sahinur Rahman Laskar[1], Bhagaban Swain[2]
[1,2]Department of Computer Science & Engineering
Assam University, Silchar, India

Abstract - In the field of information processing like data mining, information retrieval, natural language processing and machine learning, Quantum Computing play vital role for extracting the implicit, potentially useful and previously unknown information from huge sets of data. In [1] and [2], proposed two techniques of document ranking and document clustering with Quantum concept. The Quantum Clustering (QC) technique used for information processing which is basically depends on time independent Schrödinger equation for clustering the data and the Dynamic Quantum Clustering (DQC) came into existence when dynamic of the system is computed by means of the time dependent Schrödinger equation. The Dynamic Quantum Clustering (DQC),is a recent clustering technique based on physical perception from quantum mechanics where clusters are computed by means of the time dependent Schrödinger equation and clusters are identified as the minima of the potential function of the Schrödinger equation. In this paper, proposed a novel approach i.e. Pseudo Dynamic Quantum Clustering by considering time dependent Schrödinger equation for clustering the documents such that which provides an agreeable performance in terms of the quality of clusters and the efficiency of the computation in comparison of existing classical approach and earlier proposed approach [2].

Key Words: Clustering; Document Clustering; QC; DQC; Pseudo Dynamic Quantum Clustering.

1.INTRODUCTION

The objective of Clustering technique [3],[4] of automatically classify and organizing a large data collection into a finite and discrete set of data rather than provide an accurate characterization of unobserved samples generated from the same probability distribution. This problem is implied in the sense that any given set of objects can be clustered in different ways with no clear criterion for preferring one clustering over another. This challenge influenced researchers to improve classical clustering algorithm and thus Quantum Clustering (QC) technique [5],[6],[7] was developed. The QC is a broad process in the field of Quantum Information Processing (QIP), this approach brings out Schrödinger equation in existence for clustering the data but not on documents. In [2], proposed a technique which mainly focused on document clustering using time independent QC algorithm. The Dynamic Quantum Clustering (DQC) is a recent clustering method based on time dependent Schrödinger equation. This scenario is discussed in the section 3. The DQC and QC are differentiated based on the fact that DQC is computed in dynamic of the system rather than static system. This paper is the extension of the worked [2] i.e. analyses of document clustering using the proposed new approach i.e. Pseudo Dynamic Quantum Clustering.

The remainder of the paper is organized as follows. In section 2, a brief description of document clustering concept is presented. Section 3 addresses the method of dynamic quantum clustering based on time dependent Schrödinger equation. The proposed method i.e. Pseudo Dynamic Quantum Clustering is described in section 4. Experiments and results over a standard data set and comparison with existing method is given in section 5 and Finally concluding remarks for future research direction in section 6.

2. DOCUMENT CLUSTERING

Document clustering [8], [9] organizes documents into different groups called as clusters, where the documents in each cluster share some common features according to defined similarity measure. Document Clustering is different than document classification based on the fact that in document classification, the classes (and their properties) are known a priori, and documents are assigned to these classes; whereas, in document clustering, the number, properties, or membership (composition) of classes is not known in advance. Thus, classification is an example of supervised machine learning and clustering that of unsupervised machine learning. The well-known existing classical clustering techniques [10] like k-means, Agglomerative hierarchical clustering are available but limited in terms of quality of clusters which leads to the motivation of the development of quantum inspired clustering technique, where clusters are computed through the minima of the potential function of the Schrödinger equation. This scenario will be discussed in the next section.

3. DYNAMIC QUANTUM CLUSTERING

Dynamic Quantum Clustering (DQC) method [11], [12] is used for finding clusters data which is mapped into a task of quantum mechanics. The basic idea of this mapping is the analogy between each data point (i.e. a document) and a particle that is part of a quantum system and has a specific field around its location. The state of the system is represented by a function that depends on the coordinates of the particle in a specific point in time. The activation field in a location is calculated as shown in equation (1) [13].

$$\psi(x) = \sum_{i=1}^{n} e^{-(x-x_i)^2/2\sigma^2} \qquad (1)$$

Where n is the number of particles of the system and is a scale parameter. Equation (1) is also called as Parzen window estimator (or kernel density estimator), which is used for estimating the probability density of a random variable. DQC uses Parzen window estimator indirectly to construct a function whose minima are related to the clusters found by the estimator. DQC identifies local minima by letting the particles of the quantum system to "be attracted" by the local minima of the potential function. This is performed by evolution of state the system as shown in equation (2) [12]:

$$\psi(x,t) = e^{-iHt}\psi(x) \qquad (2)$$

Where H is the Hamiltonian operator, $i$is the imaginary unit and $e^{-iHt}$is the time development operator. This time evolved state is the solution to the time dependent Schrödinger equation as shown in equation (3) [11] [12].

$$i\frac{\delta\psi(x,t)}{\delta t} = H\psi(x,t) = \left\{-\frac{\nabla^2}{2m} + V(x)\right\}\psi(x,t) \qquad (3)$$

Where $-\frac{\nabla^2}{2m}$ is the kinetic energy operator, $V(x)$ the potential energy at position $x$as shown in equation (4) and the mass of the particle m is usually set equal to $\frac{1}{\sigma^2}$.

$$V(x) = \frac{E + (\frac{\delta^2}{2})\nabla^2\psi}{\psi} \qquad (4)$$

4. PSEUDO DYNAMIC QUANTUM CLUSTERING

Motivated by experience with Dynamic Quantum Clustering (DQC) technique and the task of document clustering, proposed Pseudo Dynamic Quantum Clustering approach which is implemented in two phases by considering Time dependent Schrödinger equation (3) at different energy level. The phases are as follows:

Phase 1: The Documents are clustered at ground state (i.e. the total energy of system becomes H=E=0).At H=0, the equation (2) becomes

$$\psi(x,t) = \psi(x)\,,\ [since, \quad e^{-iHt} = e^{-i.0.t} = e^0 = 1]$$

So the time dependent Schrödinger equation (3) becomes the time independent Schrödinger equation as shown in equation (5) [2], [6] at ground state (i.e. nothing but Quantum Clustering).

$$i\frac{\delta\psi(x)}{\delta t} = H\psi(x) = \left\{-\frac{\nabla^2}{2m} + V(x)\right\}\psi(x) \qquad (5)$$

The equation (1) is used to know the numbers of clusters that will produce after applying the complete process. Then local minima V(x) are computed to find the clusters centre's using equation (4). In this work for clustering it can't be known or give the number of clusters prior to the clustering process. The number of cluster is depending upon the value assigned by assumed a formula (6).

$$q = \frac{1}{2 * \sigma^2} \qquad (6)$$

Here, cluster depending variable is 'q', As per the q value decreases numbers of clusters decreases and as per the q value increases the numbers of clusters increases. The q value depends upon the sigma value assigned. The intermediate output provided, the position of document cluster points with respect to the local minima described the centres of the clusters at the respective cluster width value. Thus in this phase Pseudo Dynamic Quantum Clustering approach is transformed into so called Quantum Clustering method [2],[6].

Phase 2: Then in this phase the energy (H=E) of the system is raised from 0 to 0.1 by considering the equation (2).Here H=E=0.1 is slightly increased value is chosen to simplify the mathematical representation and considered pseudo time instances to validate the equation (2) at each of the random σ value is taken by using equation (6) to decide how many cluster number is formed after local minima $V(x)$ are computed to find the clusters centres using equation (4). By considering the point that, [12] [13] DQC identifies local minima by letting the particles (i.e.

Documents) of the quantum system to be attracted by the local minima of the potential function. This is performed by using equation (2), defined the evolution of the system. The final out put shows document cluster points at the respective cluster width σ value at the pseudo time instance using the proposed Pseudo Dynamic Quantum Clustering approach.

5. EXPERIMENTS AND RESULTS

The objective of experiments were to analyse the comparison between proposed Pseudo Dynamic Quantum Clustering approach and existing classical K-means clustering algorithm and also with earlier proposed approach with the standard evaluation measurement technique in the task of document clustering [10]. First of all described experimental setup i.e. explain source of data set used and environmental setup of the experiments. The experimental steps were performed in five steps and a pre-processing step. The pre-processing step used for representation of documents in dataset as document term matrix. Then first step used for the purpose of quantum representation of the document term matrix. The second step performed SVD to reduce the number of features considered and normalized to complete the representation process. Distribution of the document matrix in Hilbert space using Schrodinger equation were performed by the third step. The Step four computed for evolution of the system by considering pseudo time instance. Finally, the fifth step used for the computation of the local minima to find the cluster centres of documents. Then reported results were analysed with the standard measurement technique.

A. *Experiments*

Experiments were carried out on Reuters21578 (standard dataset of 8293 documents and 18933 words) in Matlab environment and reported in the results were performed using the following steps.

Pre-processing Step: Documents in dataset are pre-processed and represented as document term matrix. The term document matrix is a standard dataset of the documents.

Quantum latent semantic analysis (qLSA) used to represent document in term document matrix $TD = \{td_{ji}\}$ where $D = \{d_i\}_{i=1.....n}$ are corresponding documents and $T = \{t_j\}_{j=1.....m}$ are corresponding terms. Quantum representation of a document d_i is given by Gaussian wave function φ_i defined by:

$$\varphi_i(j) = \sqrt{\frac{td_{ji}}{\sum_{j=1}^{m} td_{ji}}} \text{ for all } j = 1 \dots \dots m$$

Step 1.Term document matrix build as Gaussian wave function matrix of $\emptyset \in \mathbb{R}^{m*n}$ setting.

$$\emptyset_{ij} = \sqrt{\frac{td_{ji}}{\sum_{j=1}^{m} td_{ji}}}$$

Step 2. Performed Singular Value Decomposition (SVD) of $\emptyset = U \sum V^T$ and formed three matrices out of which the first r columns of *U* corresponding to latent semantic space principal Eigen vector of $\emptyset\emptyset^t$ are selected and then normalized.

Step 3. Distributed normalized document matrix in the Hilbert space using Gaussian wave function of Schrodinger equation (1).

Step 4. Computed evolution of the system using equation (2) with considering Pseudo time instance.

Step 5. Computed local minima *V*(*x*) are computed to find the clusters centres of documents with gradient descent method using equation (4).

The local minima described the centers of the clusters. According to the similarity measure (jaccard measure) of centre's and documents frequency of all documents would be associated with the desire clusters and clusters are formed according to the value of σ.The value of σ computed by input value of q using assumed formula (6).

B. *Results*

The results evaluated using the similarity measurement technique so called Jaccard Measure and the Efficiency of clusters. The quality of clusters can be evaluated by observing the Efficiency and Jaccard score. The result of proposed approach compared with classical *K*-means clustering algorithm and Quantum Clustering algorithm. In Table 1, 2 and Table 3 to 4 presented case 1,2,3,4 results of classical *K*-means clustering algorithm, Quantum clustering algorithm [2] and proposed Pseudo Dynamic Quantum clustering approach. Clusters number in Quantum clustering is decided by the value of σ (i.e. the cluster width). As the σ value increases the number of clusters decreases. In Table 7, 8, 9 and 10 presented the comparisons among the three approached.

TABLE -1: EFFICIENCY OF CLUSTERS AND JACCARD VALUE USING *K*-MEANS CLUSTERING ALGORITHM.

No .of clusters(*K*)	Efficiency of cluster	Jaccard Measure
10	0.2307	0.2206
9	0.2020	0.2018
7	0.2152	0.2150
5	0.3121	0.2128

Table -2: EFFICIENCY OF CLUSTERS AND JACCARD VALUE USING QUANTUM CLUSTERING (QC) ALGORITHM [2].

σ value	No. of clusters form	Efficiency of clusters	Jaccard Measure
0.456(q=2.4)	10	0.2846	0.2257
0.471(q=2.25)	9	0.2893	0.2207
0.476(q=2.10)	7	0.3011	0.2218
0.559(q=1.60)	5	0.4000	0.2227

Table -3: CASE 1. AT Q=2.4 (σ VALUE=0.456)

Pseudo Time instance (T)	No. of Clusters formed	Efficiency	Jaccard measure
0.04	10	0.2845	0.2259
0.05	10	0.2845	0.2260
0.06	10	0.2847	0.2260
0.1	10	0.2841	0.2259
0.2	10	0.2838	0.2264

Table -4: CASE 2. AT Q=2.25 (σ VALUE =0.471)

Pseudo Time instance (T)	No. of Clusters formed	Efficiency	Jaccard measure
0.05	9	0.2890	0.2212
0.06	9	0.2891	0.2212
0.07	9	0.2892	0.2213
0.08	9	0.2893	0.2215
0.09	9	0.2891	0.2215

Table -5: CASE 3.AT Q=2.1 (σ VALUE=0.4876)

Pseudo Time instance (t)	No. of Clusters formed	Efficiency	Jaccard measure
0.04	7	0.3022	0.2208
0.05	7	0.3025	0.2211
0.06	7	0.3027	0.2214
0.1	7	0.3021	0.2212
0.2	7	0.3019	0.2216

Table -6: Case 4. At q=1.60 (σ value-0.559)

Pseudo Time instance (t)	No. of Clusters formed	Efficiency	Jaccard measure
0.01	5	0.4001	0.2230
0.04	5	0.4003	0.2242
0.05	5	0.3999	0.2243
0.1	5	0.3981	0.2242
0.2	5	0.3937	0.2239

Table -7: COMPARISON AMONG THREE APPROACHED

AT 10 NO. OF CLUSTERS FORMED

At 10 No. of Clusters formed	*K*-means clustering	Quantum Clustering (QC) σ =0.456 (q=2.4)	Pseudo Dynamic Quantum Clustering σ =0.456 (q=2.4)
Pseudo time instance (t)			0.06
Efficiency	0.2307	0.2846	0.2847
Jaccard measure	0.2206	0.2257	0.2260

Table -8: COMPARISON AMONG THREE APPROACHED

AT 9 NO. OF CLUSTERS FORMED

At 9 No. of Clusters formed	*K*-means clustering	Quantum Clustering (QC) σ =0.471 (q=2.25)	Pseudo Dynamic Quantum Clustering σ =0.471 (q=2.25)
Pseudo Time instance (t)			0.08
Efficiency	0.2020	0.2893	0.2893
Jaccard measure	0.2018	0.2207	0.2215

Table -9: COMPARISON AMONG THREE APPROACHED

AT 7 NO. OF CLUSTERS FORMED

At 7 No. of Clusters formed	*K*-means clustering	Quantum Clustering (QC) σ =0.4876 (q=2.10)	Pseudo Dynamic Quantum Clustering σ =0.4876 (q=2.10)
Pseudo Time instance (t)			0.06
Efficiency	0.2152	0.3011	0.3027
Jaccard measure	0.2150	0.2208	0.2214

Table -10: COMPARISON AMONG THREE APPROACHED

At 5 N0. of clusters formed

At 5 No. of Clusters formed	*K*-means clustering	Quantum Clustering (QC) σ =0.559 (q=0.16)	Pseudo Dynamic Quantum Clustering σ =0.559 (q=0.16)
Pseudo Time instance (t)			0.04
Efficiency	0.3121	0.4000	0.4003
Jaccard measure	0.2128	0.2227	0.2242

Table VII, VIII, IX and X, it was analyzed results through the comparisons at 10, 9, 7 and 5 no. of clusters formed. At 10 No. of Clusters formed, Jaccard measure and efficiency of existing K-means clustering algorithm are 0.2206, 2307 Quantum Clustering (QC) [2] algorithm are 0.2257,2846 and proposed Pseudo Dynamic Quantum Clustering are 0.2260,2847 at Pseudo time instance (t) = 0.06, Similarly at 9, 7 and 5 no. of clusters formed, it is observed that the proposed Pseudo Dynamic Quantum Clustering provides better quality of clusters (Jaccard measure) and efficiency than Quantum Clustering (QC) algorithm [2] and existing K-means clustering algorithm. Fig. 1 below shows the sample output of proposed Pseudo Dynamic Quantum Clustering approach.

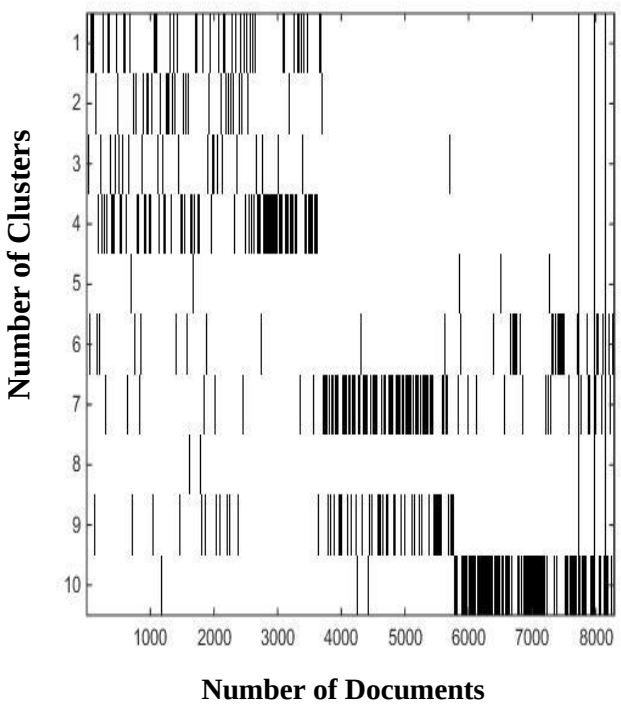

Fig.1. Pseudo Dynamic Quantum Clustering at 10 no. of clusters where pseudo time instance (t)=0.06.

6. FINAL REMARKS

In this paper, it was analyzed document clustering using proposed Pseudo Dynamic Quantum Clustering approach which provided an agreeable performance in terms of the quality of clusters and the efficiency of the computation. This work presented a novel approach which partially applied DQC algorithm by assumed Pseudo time instance in case of standard document data set where the features of the data not directly related to time. If the features of data directly related to time like sequence of images or frames in case of dynamic video or a combination of text, audio and images with a time dimension which is an important form of multimedia information where features directly relate to time then the time dependent Quantum Clustering technique i.e. DQC concept can be applied. In the domain of multimedia information retrieval, clustering techniques are basically used in the task of video segmentation. The video segmentation plays an important role in video indexing and retrieval, a wide spectrum of promising applications, motivating the interest of researchers worldwide. This encourages into the future research. So the future direction for continuing the research is the implementation of clustering approach with quantum concept for video segmentation in the area of video indexing and retrieval.

REFERENCES

[1] Laskar, Sahinur Rahman; Swain, Bhagaban "Analyzing Quantum Probability Ranking Principle with the Concept of Hyperspace Analogue To Language (HAL)", in Advanced Communicating and Communication(ISACC), 2015 International Symposium on, vol.,no.,pp.266-271, 14-15 Sept. (2015).

[2] R. Bhagawati, S. R. Laskar, B. Swain, "Document Clustering using Quantum Clustering Algorithm", 2016 International Conference Microelectronics, Computing and Communications (MicroCom), pp. 1-4, 23-25 Jan, (2016).

[3] K.Kameshwaran and K.Malarvizhi, "Survey on Clustering Techniques in D[]ata Mining" International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 5 (2) , 2272-2276 (2014).

[4] Xu, R., Ii: Survey of clustering algorithms. IEEE Transactions on Neural Networks 16(3), 645–678 (2005).

[5] David Horn and Assaf Gottlieb. Quantum Clustering.School of Physics and Astronomy Raymond and Beverly Sackler Faculty of Exact Sciences.Tel Aviv University, Tel Aviv 69978, Israel (2008).

[6] Esma A., Gilles Brassard and S. Gambs,"Quantum Clustering Algorithms", In the Proceedings of the 24th International Conference on Machine Learning, ICML (2007).

[7] Horn, D. & Gottlieb, A. The method of quantum clustering. Proceedings of the Neural Information Processing Systems: NIPS'01 (pp. 769–776) (2001).

[8] Neepa Shah and Sunita Mahajan. Document Clustering: A Detailed Review, International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 4– No.5,October (2012).

[9] A. Huang, "Similarity measures for text document clustering," In Proc. of the Sixth New Zealand Computer Science Research Student Conference NZCSRSC, pp. 49—56 (2008).

[10] Michael Steinbach, George Karypis, Vipin Kumar. A Comparison of Document Clustering Techniques. Department of Computer Science and Engineering,, University of Minnesota. Technical Report #00-034 (2000).

[11] Weinstein, M., Horn, D.: Dynamic quantum clustering: A method for visual exploration of structures in data. Phys. Rev. E 80(6), 066117 (2009).

[12] E. Di Buccio and G. Di Nunzio. Distilling relevant documents by means of dynamic quantum clustering. In Proceedings of ICTIR-11, 3rd International Conference on the Theory of Information Retrieval, pages 360{363, Bertinoro, Italy, September (2011).

[13] E. Di Buccio and G. Di Nunzio. Envisioning Dynamic Quantum Clustering in Information Retrieval D. Song et al. (Eds.): QI 2011, LNCS 7052, pp. 211–216, 2011.Springer-Verlag Berlin Heidelberg (2011).