IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ADAPTIVE SELF PACED DEEP CLUSTERING WITH DATA AUGUMENTION

VINNAKOTA GOWTHAMI PRIYA [#1], K.VENKATESH [#2]

[#1] MSC Student, Master of Computer Science,

D.N.R. College, P.G.Courses & Research Center, Bhimavaram, AP, India.

[#2] Assistant Professor, Master of Computer Applications,

D.N.R. College, P.G.Courses & Research Center, Bhimavaram, AP, India.

Abstract

Deep clustering is one of the most prominent clustering techniques compared with superior clustering techniques by jointly performing feature learning and cluster assignment. Although numerous deep clustering algorithms have emerged in various applications, most of them fail to learn robust cluster-oriented features which in turn hurt the final clustering performance. In order to overcome all these problems, we try to propose a two-stage deep clustering algorithm by incorporating data augmentation and self-paced learning. In the first stage, we try to train the system with all the inputs which are required for identifying the post with category name and rank. In the next stage we just want to identify the test condition in which user try to recommend the post to others and combine a post based on category.

1. INTRODUCTION

CLUSTERING has been intensively studied in the data mining and machine learning community. Conventional clustering algorithms such as k-means [1], Gaussian Mixture Model (GMM) [2], and hierarchical clustering [3] generally group data on handcrafted features according to intrinsic similarity. It is well known that these features are designed for general purpose and may not be suitable for a specific task. Some clustering algorithms, including spectral clustering (SC) [4] and kernel k-means [5], [6], transform data into a new feature space in which the clustering task becomes much easier.

The resultant feature is task-specific, i.e., more suitable for clustering. In this paper, we refer to as Deep Clustering the algorithm that jointly performs feature learning and clustering by DNNs. Most existing deep clustering algorithms tune the parameters of the DNN by using a loss function defined by the cluster centers and assignments, which are generally obtained based on the outputs of the DNN in the last iteration. However, we observe that these methods do not explicitly consider the effect of marginal examples on the network training. As the goal of the DNN is to learn features that are more suitable for clustering, the examples near cluster boundaries may not provide convincing guidance. This is in contrast to supervised learning where all target labels are given beforehand and hence all examples can give trustworthy supervisory signals. Actually, marginal examples in supervised learning play a more important role in searching class boundaries.

We empirically validate that, in deep clustering, unreliable examples near cluster boundaries could confuse or even mislead the training process of the DNN, leading to unsatisfying performance. On the other hand, these clustering algorithms also overlook the technique of data augmentation which has been widely employed in supervised deep learning models to improve the generalization. Neglecting these two ingredients leads to the failure of leaning robust cluster-oriented features. In this paper, we propose an Adaptive Self-Paced deep Clustering with Data Augmentation (ASPC-DA) algorithm which is comprised of two stages: pre training and fine tuning. In the stage of pre training, we train an auto encoder with augmented data by minimizing the reconstruction loss. As we know, the auto encoder can transform data from relatively high dimensional and sparse space to low dimensional and compact representation space. We use augmented data to enhance the smoothness of the manifold on which the learned representations lie.

Then in the second stage, we fine tune the encoder (feature extractor) by using a clustering loss defined as the within-cluster sum of squares. To stabilize the training process, we adopt adaptive self-paced learning to select "easy" (near cluster centers) examples as training set and gradually add harder examples. Different from typical self-paced learning, our adaptive variant is free of hyper-parameters and always keeps marginal examples out of training. We also adopt the same type of data augmentation as in the pre training stage to further facilitate the feature learning. Our experiments show a great competitive edge over state-of-the-art clustering algorithms on various datasets. The main contributions of this paper are highlighted as follows: To learn robust cluster-oriented features, we propose a simple but effective deep clustering model that incorporates self paced learning and data augmentation.

PROBLEM STATEMENT

However, these methods generally have a limited capacity for transformation or suffer from high computational complexity. With the development of deep learning, Deep Neural Networks (DNNs) have shown amazing power of highly nonlinear transformation (or feature learning). Recently, some researches [7], [8], [9], [10], [11], [12] adopt DNNs to perform clustering, showing dramatic improvement on clustering performance. The basic idea is that good feature helps produce good clustering result, and the latter in turn guides DNNs to learn the better feature.

PURPOSE

To the best of our knowledge, this is the first work to introduce these two well-studied techniques in supervised learning into unsupervised deep clustering field. We derive an adaptive self-paced learning algorithm without extra hyper-parameter. By using the adaptive self-paced learning, our model can eliminate the negative effect of the examples near cluster boundaries in the process of feature learning. Extensive experiments are conducted and the results validate the effectiveness of our ASPC-DA algorithm.

OBJECTIVE

Deep clustering is one of the most prominent clustering techniques compared with superior clustering techniques by jointly performing feature learning and cluster assignment. Although numerous deep clustering algorithms have emerged in various applications, most of them fail to learn robust cluster-oriented features which in turn hurt the final clustering performance.

1.4 SCOPE

. In order to overcome all these problems, we try to propose a two-stage deep clustering algorithm by incorporating data augmentation and self-paced learning. In the first stage, we try to train the system with all the inputs which are required for identifying the post with category name and rank. In the next stage we just want to identify the test condition in which user try to recommend the post to others and combine a post based on category.

2. LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool, it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, ten next steps are to determine which operating system and language used for developing the tool. Once the programmers start building the tool, the programmers need lot of external support. This support obtained from senior programmers, from book or from websites. Before building the system the above consideration r taken into for developing the proposed system.

1. Social contextual recommendation

Authors: M. Jiang, P. Cui, R. Liu, Q. Yang, F. Wang, W. Zhu, and S. Yang

Exponential growth of information generated by online social networks demands effective and scalable recommender systems to give useful results. Traditional techniques become unqualified because they ignore social relation data; existing social recommendation approaches consider social network structure, but social contextual information has not been fully considered. It is significant and challenging to fuse social contextual factors which are derived from users' motivation of social behaviors into social recommendation. In this paper, we investigate the social recommendation problem on the basis of psychology and sociology studies, which exhibit two important factors: individual preference and interpersonal influence. We first present the particular importance of these two factors in online behavior prediction. Then we propose a novel probabilistic matrix factorization method to fuse them in latent space. We further provide a scalable algorithm which can incrementally process the large scale data. We conduct experiments on both Facebook style bidirectional and Twitter style unidirectional social network data sets. The

empirical results and analysis on these two large data sets demonstrate that our method significantly outperforms the existing approaches.approaches.

2. Personalized recommendation based on reviews and ratings alleviating the sparsity problem of collaborative filtering

Authors: J. Xu, X. Zheng, W. Ding

With the development of e-commerce, shopping on-line is becoming more and more popular. When we need to decide whether to purchase a product or not on line, the opinions of others become important. The convenience of new web technologies enables us to freely express our opinions and reviews for various products we have purchased which leads to a serious problem, information overloading. How to mine these review data to understand customers' preferences and make recommendations is crucial to merchants and researchers. Traditional collaborative filtering (CF) algorithm is one of the most successful recommendation system technologies. The core idea of CF algorithm is to recommend products based on other people who have similar tastes with target users. However, the ability of CF is limited by the sparsity problem, which is very common in reality. The reason derives from the fact that traditional CF method only takes users' ratings into account. In this paper, we propose a new personalized recommendation model, i.e. topic model based collaborative filtering (TMCF) utilizing users' reviews and ratings. We exploit extended LDA model to generate topic allocations for each review and then obtain each user's preference. Moreover, a new metric is designed to measure similarity between users alleviating the sparsity problem to a large extent. Finally, recommendations are made based on similar users' ratings. Experiments on seven data sets indicate better prediction accuracy than other traditional and state-of-the-art methods with substantial improvement in alleviating the sparsity problem.

3. Semantic-based location recommendation with multimodal venue semantics

Authors: X. Wang, Y. Zhao, L. Nie, Y. Gao

In recent years, we have witnessed a flourishing of location -based social networks. A well-formed representation of location knowledge is desired to cater to the need of location sensing, browsing, navigation and querying. In this paper, we aim to study the semantics of point-of-interest (POI) by exploiting the abundant heterogeneous user generated content (UGC) from different social networks. Our idea is to explore the text descriptions, photos, user check-in patterns, and venue context for location semantic similarity measurement. We argue that the venue semantics play an important role in user check-in behavior. Based on this argument, a unified POI recommendation algorithm is proposed by incorporating venue semantics as a regularizer. In addition to deriving user preference based on user-venue check-in information, we place special emphasis on location semantic similarity. Finally, we conduct a comprehensive performance evaluation of location semantic similarity and location recommendation over a real world dataset collected from Foursquare and Instagram. Experimental results show that the UGC information can well characterize the venue semantics, which help to improve the recommendation performance.

4. Review expert collaborative recommendation algorithm based on topic relationship

Authors: Shengxiang Gao, Zhengtao Yu, Linbin Shi

The project review information plays an important role in the recommendation of review experts. In this paper, we aim to determine review expert's rating by using the historical rating records and the final decision results on the previous projects, and by means of some rules, we construct a rating matrix for projects and experts. For the data sparseness

problem of the rating matrix and the "cold start" problem of new expert recommendation, we assume that those projects/experts with similar topics have similar feature vectors and propose a review expert collaborative recommendation algorithm based on topic relationship. Firstly, we obtain topics of projects/experts based on latent Dirichlet allocation (LDA) model, and build the topic relationship network of projects/experts. Then, through the topic relationship between projects/experts, we find a neighbor collection which shares the largest similarity with target project/expert, and integrate the collection into the collaborative filtering recommendation algorithm based on matrix factorization. Finally, by learning the rating matrix to get feature vectors of the projects and experts, we can predict the ratings that a target project will give candidate review experts, and thus achieve the review expert recommendation. Experiments on real data set show that the proposed method could predict the review expert rating more effectively, and improve the recommendation effect of review experts.

3. EXISTING SYSTEM

In the existing system, there was no method like sentiment-based rating prediction method (RPS) method to improve prediction accuracy in recommender systems .In the existing system the analysis can be conducted on three different levels:

i. review-level,
ii. sentence-level, and
iii. phrase-level.

In the existing system there is no facility to automatically classify the messages and recommend to others.

LIMITATION OF EXISTING SYSTEM

The following are the main limitations of the existing system. They are as follows:

1) The existing work mainly focuses on classifying users into binary Review system(i.e. positive or negative), and they do not go further in mining user's sentiment.

2) In the existing system, we cant able to separate the reviews automatically .

3) In the existing system, automatic recommendation is not possible.

4. PROPOSED SYSTEM

We try to propose a two-stage deep clustering algorithm by incorporating data augmentation and self-paced learning. In the first stage, we try to train the system with all the inputs which are required for identifying the post with category name and rank. In the next stage we just want to identify the test condition in which user try to recommend the post to others and combine a post based on category.

ADVANTAGES OF THE PROPOSED SYSTEM

The following are the advantages of the proposed system, they are as follows:

1. Here we can do deep clustering algorithm by incorporating data augmentation and self-paced learning
2. The purpose of our proposed approach is very accurate to predict the model.

This is purely used for recommendation to others.

5. SOFTWARE PROJECT MODULES

Implementation is the stage where the theoretical design is converted into programmatically manner. In this stage we will divide the application into a number of modules and then coded for deployment. The front end of the application takes JSP,HTML and Java Beans and as a Back-End Data base we took My SQL data base. The proposed application is divided into mainly 2 modules and they are many sub-modules present in these main modules. Now let us look about them in detail as follows:

1) Admin Module
2) User Module

MODULES DESCRIPTION:

5.1 ADMIN MODULE

In this module, the Admin has to login by using valid user name and password. After login successful he can do some operations such as Add Domain ,Add Posts based on Domain ,List All Posts with ranks ,List All Recommended Posts based on Domain ,List All reviewed Posts ,List Users and authorize ,List All Search History ,View Similar Domain users based on Domain sign, View Similar user services

Search Transactions

This is controlled by admin; the admin can view the search history details. If he clicks on search history link, it will show the list of searched user details with their tags such as user name, searched user, time and date.

Request & Response

In this module, the admin can view the all the friend request and response. Here all the request and response will be stored with their tags such as Id, requested user photo, requested user name, user name request to, status and time & date. If the user accepts the request then status is accepted or else the status is waiting.

5.2 USER MODULE

In this module, there are n numbers of users are present. User should register before doing some operations. And register user details are stored in user module. After registration successful he has to login by using authorized user name and password. Login successful he will do some operations like View your Details & Search users based on Domain sign ,Search for Posts & View specified post and recommend to other by feeding your interest on the Posts, View my search History ,View recommends based on Domain ,View user interests on the post.

Search Users

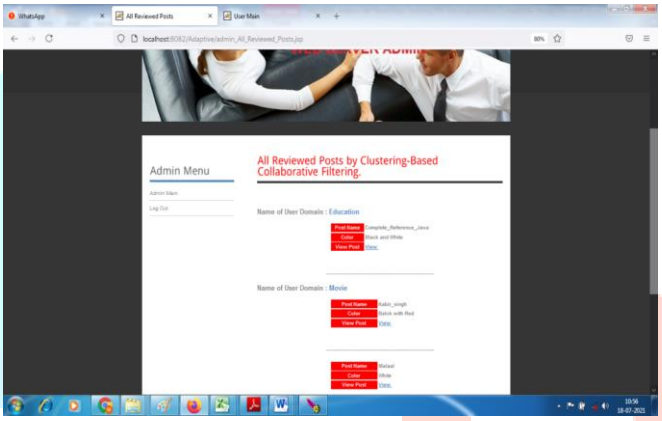
The user can search the users based on users and the server will give response to the user like User name, user image, E mail id, phone number and date of birth. If you want send friend request to particular receiver then click on follow, then request will send to the user.

Followers

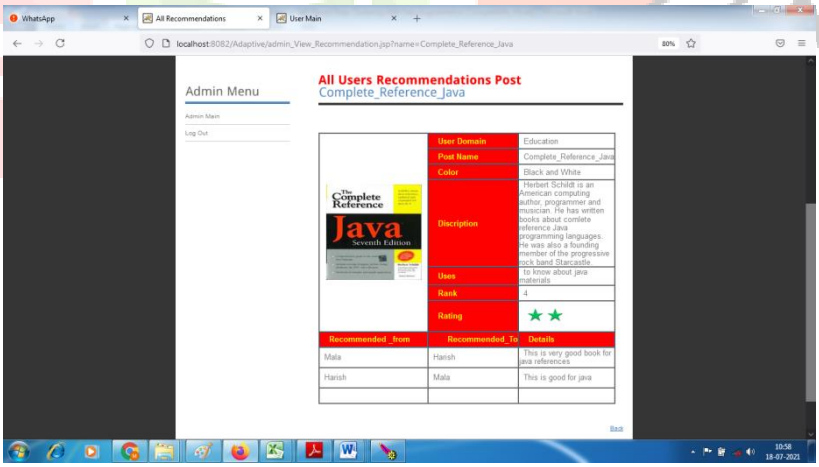
In this module, we can view the followers' details with their tags such as user name, user image, date of birth, E mail ID, phone number and ranks.

6. OUTPUT RESULTS

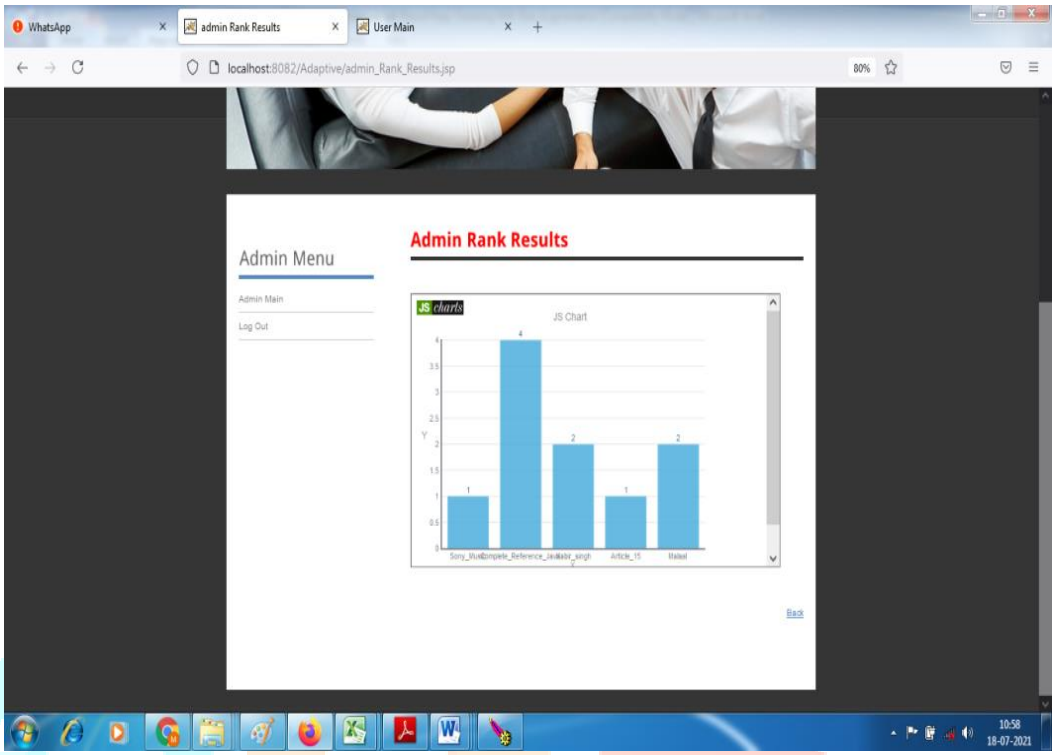
6.1 Admin Views All Reviewed Post



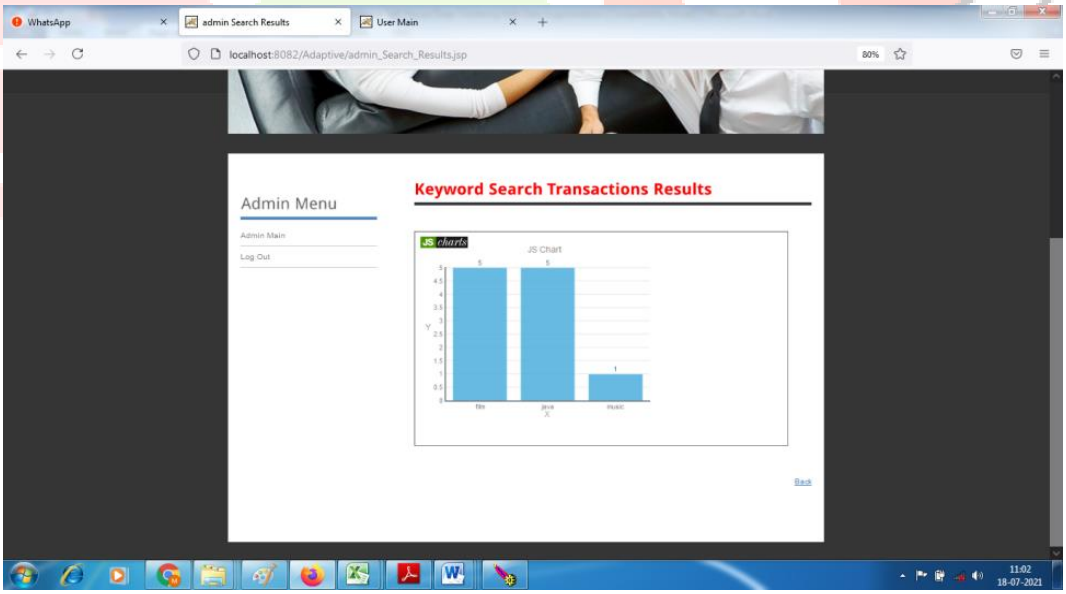
6. 2 Admin Views Recommendation Post



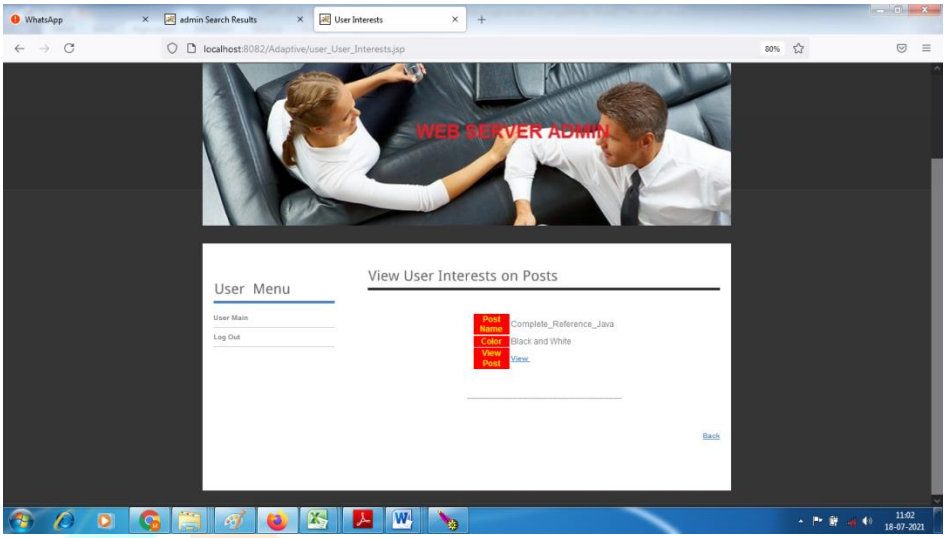
6.3 Admin Can View Rank of All Post

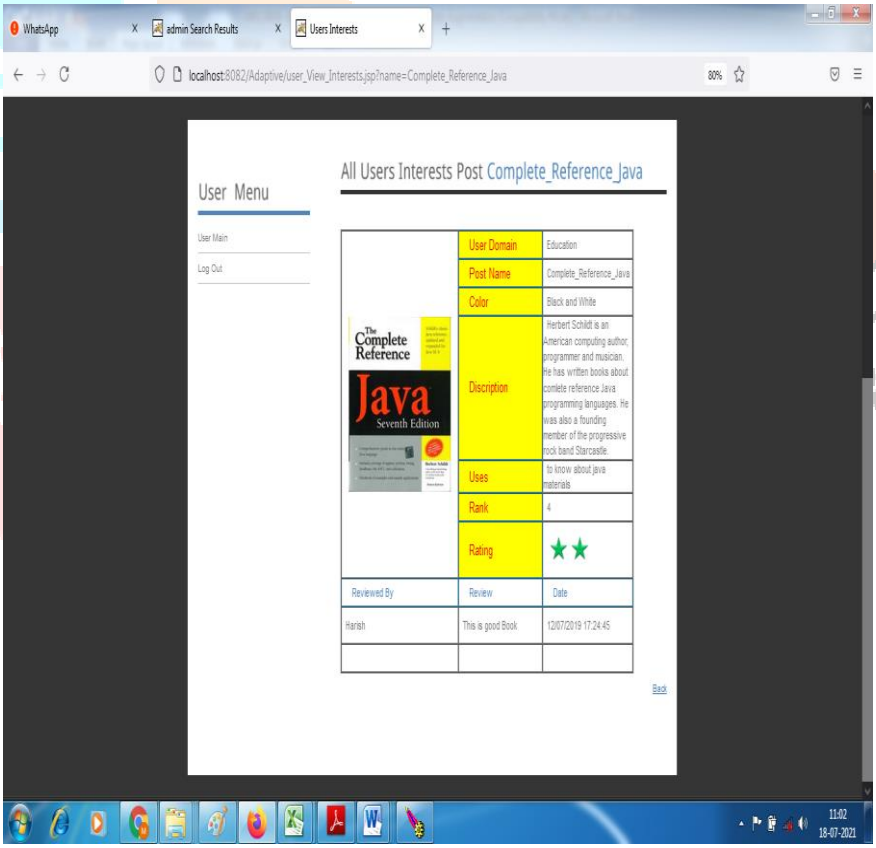


6.4 Admin Can View Search results in rank chart

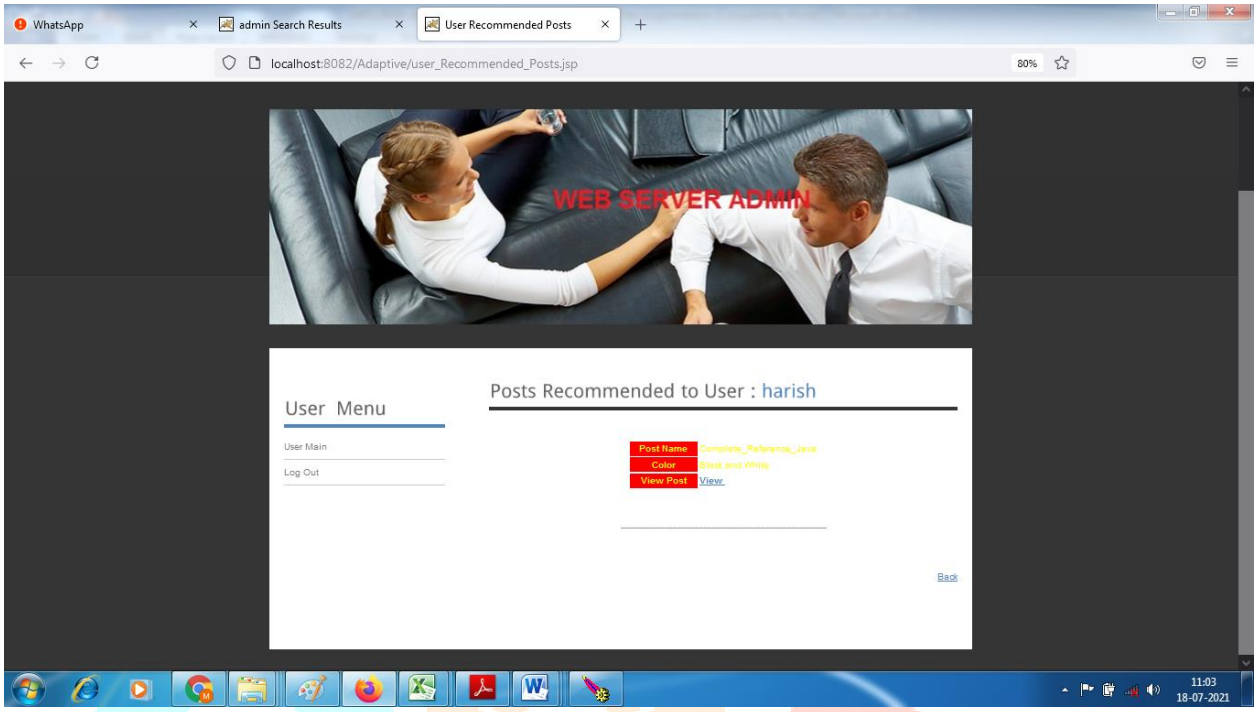


6.5 User Can View User interest on Post





6.6 User can View the Recommend Post



7. CONCLUSION

We proposed an Adaptive Self-Paced deep Clustering with Data Augmentation (ASPC-DA) algorithm to learn robust cluster-oriented features. Our ASPC-DA excludes the examples near cluster boundaries from training by gradually adding "easy" (close to cluster centers) examples. We formulated the process of selecting examples and proposed an adaptive self-paced learning algorithm which does not introduce extra hyper-parameters

8. REFERENCES

[1] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, no. 14. Oakland, CA, USA., 1967, pp. 281–297.

[2] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.

[3] A. K. Jain, "Data clustering: 50 years beyond k-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, 2010.

[4] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in Proceedings of Advances in Neural Information Processing Systems (NIPS), 2002, pp. 849–856.

[5] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: spectral clustering and normalized cuts," in Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,
2004.

[6] X. Liu, Y. Dou, J. Yin, L. Wang, and E. Zhu, "Multiple kernel kmeans clustering with matrix-induced regularization," in Proceedings of AAAI Conference on Artificial Intelligence (AAAI), 2016, pp. 1888–1894.

[7] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in Proceedings of International Conference on Machine Learning (ICML), 2016, pp. 478–487.

[8] X. Peng, S. Xiao, J. Feng, W.-Y. Yau, and Z. Yi, "Deep subspace clustering with sparsity prior." in Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), 2016, pp. 1925–1931.

[9] X. Peng, J. Feng, J. Lu, W.-Y. Yau, and Z. Yi, "Cascade subspace clustering," in Proceedings of AAAI Conference on Artificial Intelligence (AAAI), 2017, pp. 2478–2484.

[10] K. G. Dizaji, A. Herandi, and H. Huang, "Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization," arXiv preprint arXiv:1704.06327, 2017.

[11] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation," in Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), 2017, pp. 1753–1759.

[12] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong, "Towards kmeans- friendly spaces: Simultaneous deep learning and clustering," in Proceedings of International Conference on Machine Learning (ICML), vol. 70, 2017, pp. 3861–3870.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in Proceedings of International Conference on Learning Representations (ICLR), 2014.