**CSE3024- Web Mining**

***Digital Assignment - I***

***<Project Title>***

*By*

21X XX1001 Sandeep J

21MCS1012 Pavithraa S

21MCS1019 Sirigiri Gayathri

B.Tech CSE with Specializaton

*Submitted to*

**Dr.A.Bhuvaneswari,**

Assistant Professor Senior,

SCOPE, VIT, Chennai

**School of Computer Science and Engineering**

****

*January 2023*

*­­­*

**School of Computing Science and Engineering**

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

WINTER SEM 22-23

**Worklet details**

|  |  |  |
| --- | --- | --- |
| Programme | B.Tech with Specialization | |
| Course Name / Code |  | |
| Slot |  | |
| Faculty Name |  | |
| J Component Title |  | |
| Team Members Name | Reg. No |  |  |
|  |  |
|  |  |
|  |  |

**Team Members(s) Contributions – Tentatively planned for implementation:**

|  |  |
| --- | --- |
| *Worklet Tasks* | *Contributor’s Names* |
| Database connection and integration using Pymongo |  |
| Preprocessing |  |
| Model building |  |
| Visualization |  |
| Technical Report writing |  |
| Presentation preparation |  |

**ABSTRACT**

*<150-350 Words>*

1. Introduction

*<description of problem statement, objectives, challenges of your project) 200 to 300 words>*

1. Literature Survey (sample)

*<Minimum 10 papers should be surveyed>*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl no** | **Title** | **Author / Journal name / Year** | **Technique** | **Result** |
| 1 | Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks |  [C. Okan Sakar](https://link.springer.com/article/10.1007/s00521-018-3523-0#auth-C__Okan-Sakar),   [S. Olcay Polat](https://link.springer.com/article/10.1007/s00521-018-3523-0#auth-S__Olcay-Polat),   [Mete](https://link.springer.com/article/10.1007/s00521-018-3523-0" \l "auth-Mete-Katircioglu)  [Katircioglu](https://link.springer.com/article/10.1007/s00521-018-3523-0" \l "auth-Mete-Katircioglu) &   [Yomi Kastro](https://link.springer.com/article/10.1007/s00521-018-3523-0#auth-Yomi-Kastro)  Springer  2018 | RF,SVM,MLP | Best result was given by MLP |
| 2 | Algorithms for Clustering ClickStream Data | Panagiotis Antonellis, Christos Makris, Nikos Tsirakis  Information Processing Letters, 2009 - Elsevier | Model for clustering clickstream data which applies three different phases in the data processing, and is validated through a set of experiments.  KNN, meta clustering procedure | By combining the initial merged table with the data meta-clustering results it provide s recommendations to users in order to visit other unvisited pages according to the moves of users that all belong to the same cluster. The meta-clustering procedure gives an overview of user’s behaviour and this helps us recognize the useful clusters that have users who have stable preferences about page content. |
|  |  |  |  |  |
| 3 | Mining and Analysis of Clickstream Patterns | H. Hannah Inbarani and K. Thangavel  Springer 2009 | A partitional algorithm namely Multi Pass Combined Standard Deviation(CSD) Means algorithm which automatically generates the optimum number of clusters from the web clickstream patterns. The quality of clusters obtained using these algorithms are compared using K-Means algorithm, Rough K-Means algorithm and model based algorithms ANTCLUST and ACCANTCLUST | A comparative analysis is made with K-Means algorithm, Rough K-Means and Model based algorithms ANTCLUST and ACCANTCLUST. The study of validity measure shows the effectiveness of the proposed algorithm for mining web clickstream patterns. Experimental results also show the stability and accuracy of the proposed algorithm. Though ANTCLUST and ACCANTCLUST estimates the number of clusters from the data set, experimental results clearly show that they are not stable and these algorithms take enormous amount of time for the convergence of clusters |
| 4 | Visualization and Analysis of Clickstream Data of Online Stores for Understanding Web Merchandising | JUHNYOUNG LEE  MARK PODLASECK EDITH SCHONBERG ROBERT HOCH  2001  Springer | The system visualizes the effectiveness of Web merchandising from two different points of view by using two different visualization techniques: visualization of sessions by using parallel coordinates and visualization of product performance by using starfield graphs | The presented system visualizes the progression of sessions in the store, i.e., the conversions from one shopping step to another, and so provides insight into the effectiveness of each step’s design. By associating the sessions with attributes that categorize them such as the referrers, host names, length, the shopping metaphors and merchandising cues, the sessions and their conversions could be subdivided. The categorization of sessions helps to understand how sessions with different category values react to the site differently |
| 5 | Representation of click stream datasequences for learning user navigational behavior by using embeddings | Erdi Olmezogullari ∗,Mehmet S. Aktas  IEEE 2020 | Application of embedding approaches to represent click-stream data sequences, to enable machine learning algorithms learn the users’ navigational behaviors on web-sites. By utilizing embedding representation, we propose an algorithm that takes clickstream data as input and creates clustered sequential patterns.  Word2Vec, Deepwalk , Node2Vec | Results indicate that the best clustering results are obtained when using the Word2Vec embedding data representation. The Word2Vec approach outperforms the other embedding data representation approaches in creating the best quality clustering results. It accurately identifies the 70 percentage of the predefined annotations that falls into actual 7 clusters as defined according to metric v-score. |
| 6 | A Deep Markov Model for Clickstream Analytics in Online Shopping | Yilmazcan Ozyurt, Ce Zhang, Tobias Hatt, Stefan Feuerriegel  ACM 2020 | Formalize a tailored attentive deep Markov model called ClickstreamDMM for predicting the risk of user exits without purchase in e-commerce web sessions. ClickstreamDMM combines an attention network to learn long-term dependencies in clickstream data and a latent variable model to capture different shopping phases | ClickstreamDMM yields a substantial performance improvement in AUROC by 11.5 % and in AUPRC by 12.7 % |
| 7 | A Web Page Prediction Model Based on Click-Stream Tree Representation of User Behavior | Sule Gunduz, M. Tamer Ozsu  ACM 2003 | Proposal of a new model that considers both the order information of pages in a session and the time spent on them. Clustering user sessions based on their pair-wise similarity and represent the resulting clusters by a click-stream tree. The new user session is then assigned to a cluster based on a similarity measure. The click-stream tree of that cluster is used to generate the recommendation set. The model can be used as part of a cache prefetching system as well as a recommendation model. | The experiments show that the model can be used on Web sites with different structures. To confirm findings, they compare their  model to three other recommendation models. Results show that the model improves the efficiency and effectiveness significantly. |
| 8 | Predicting web site audience demographics for web advertising targeting using multi-web site clickstream data | Koen W. De Bock, Dirk Van den Poel  Iospress 2010 | The methodology involves the transformation of web site visitors’ clickstream patterns to a set of features and the training of Random Forest classifiers that generate predictions for gender, age, educational level and occupation category. These demographic predictions can support online advertisement targeting (i) as an additional input in personalized advertising or behavioral targeting, in order to restrict ad targeting to demographically defined target groups, or (ii) as an input for aggregated demographic web site visitor profiles that support marketing managers in selecting web sites and achieving an optimal correspondence between target groups and web site audience composition. | The average absolute error is the highest for the single-user test sample data, but when looking at the most realistic setting, i.e., out-of-period data consisting of a mix of single and multi-user data, this average drops to 2.85 percent. Overall, this figure demonstrates the practical value of the model set to create usable demographic web site audience profiles. When looking at overall, but model specific error figures, strikingly, error figures are the highest for the gender model (4.33 percent), while for the multi-class characteristics age, occupation and education, these average errors are considerably lower (resp. 3.10, 3.87 and 2.85 percent). This is in contrast to the findings of the classification performance evaluation, which demonstrated the best results for the binary gender model. However, an explanation can be found in the fact that web sites differ more strongly in the gender distribution of their audience than in terms of the other demographic characteristics..Despite the emergence of advertisement personalization and behavioral targeting, demographic information still plays an important role for web advertising purposes. |
| 9 | Mining Significant Usage Patterns from Clickstream Data | Lin Lu, Margaret Dunham, and Yu Meng  Springer 2006 | A technique to generate Significant Usage Patterns (SUP) is proposed and used to acquire significant “user preferred navigational trails”. The technique uses pipelined processing phases including sub-abstraction of sessionized Web clickstreams, clustering of the abstracted Web sessions, concept-based abstraction of the clustered sessions, and SUP generation. Using this technique, valuable customer behavior information can be extracted by Web site practitioners. | On average purchase sessions are longer than those sessions without purchase. This illustrates that users usually request more page views when they are about to make a purchase than when they visit an online store without purchasing. This can be explained by the fact that users normally would like to review the information as well as to compare the price, the quality and etc. for the product(s) of their interest before buying them. |
| 10 | Multiple factor hierarchical clustering algorithm for large scale web page and search engine clickstream data | Gang Kou · Chunwei Lou | A hierarchical clustering method that combines multiple factors to identify clusters of web pages that can satisfy users’ information needs. The clusters are primarily envisioned to be used for search and navigation and potentially for some form of visualization as well. An experiment on Clickstream data from a processional search engine was conducted to examine the results shown that the clustering method is effective and efficient, in terms of both objective and subjective measures. | Users tend to click things that are higher on a list regardless of how good the recommendation is. the ranking of more than 90% of the items in the first 10 ranking has been changed in the clustering results and the average clickthrough precision of the new ranking by clusters is 12.90% higher than original search result ranking. This is a significant improvement since the total number of sessions is more than 20 million. This algorithm is able to efficiently assign a cluster request of web pages to its closest cluster in the presence of a large collection of text data and deal with noisy data without affecting the computation of the centroid of clusters. |

1. **Algorithms / Techniques description**

*<Pseudocode >*

1. **Github Repository Link (where your j comp project work can be seen for assessment)**

**REFERENCES**

1. Sakar, C.O., Polat, S.O., Katircioglu, M. *et al.* Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Comput & Applic* **31**, 6893–6908 (2019). https://doi.org/10.1007/s00521-018-3523-0
2. Antonellis, P., Makris, C., & Tsirakis, N. (2009). Algorithms for clustering clickstream data. *Information Processing Letters*, *109*(8), 381-385.
3. Inbarani, H. H., & Thangavel, K. (2009). Mining and analysis of clickstream patterns. *Foundations of Computational, IntelligenceVolume 6: Data Mining*, 3-27.
4. Lee, J., Podlaseck, M., Schonberg, E., & Hoch, R. (2001). Visualization and analysis of clickstream data of online stores for understanding web merchandising. *Data mining and knowledge discovery*, *5*, 59-84.
5. Olmezogullari, E., & Aktas, M. S. (2020, December). Representation of click-stream datasequences for learning user navigational behavior by using embeddings. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 3173-3179). IEEE.
6. Ozyurt, Y., Hatt, T., Zhang, C., & Feuerriegel, S. (2022, April). A deep Markov model for clickstream analytics in online shopping. In *Proceedings of the ACM Web Conference 2022* (pp. 3071-3081).
7. Gündüz, Ş., & Özsu, M. T. (2003, August). A web page prediction model based on click-stream tree representation of user behavior. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 535-540).
8. De Bock, K., & Van den Poel, D. (2010). Predicting website audience demographics forweb advertising targeting using multi-website clickstream data. *Fundamenta Informaticae*, *98*(1), 49-70.
9. Lu, L., Dunham, M., & Meng, Y. (2006). Mining significant usage patterns from clickstream data. In *Advances in Web Mining and Web Usage Analysis: 7th International Workshop on Knowledge Discovery on the Web, WebKDD 2005, Chicago, IL, USA, August 21, 2005. Revised Papers 7* (pp. 1-17). Springer Berlin Heidelberg.
10. Kou, G., & Lou, C. (2012). Multiple factor hierarchical clustering algorithm for large scale web page and search engine clickstream data. *Annals of Operations Research*, *197*(1), 123.