Improvised DBSC and transductive SVM for hybrid recommendation systems

K. Anima
Research Scholar, Department of Computer Science, AJK College of Arts and Science, Coimbatore, Tamilnadu - 641105, India
Corresponding author email: animaabhi@gmail.com

Dr P. Senthil Kumar
Assistant Professor, Department of Computer Science, AJK College of Arts and Science, Coimbatore, Tamilnadu - 641105, India
senthilkumar@ajkcas.com

Abstract---Recommendation Systems play a crucial role in improving the business strategies and providing the customers or users with the choice to opt for a product or service. Given the growth and power of information analytics tools and inclination of data mining techniques, a wide range of recommender systems particularly for the web based users have been evolved over the decade. The conventional “One model – Fit for all” model fails in most of the cases due to the common fact that not all users have same interests and have paved way for tailor made recommender systems. It is hence highly important to provide a personalized framework which can accumulate specific interests of the users. Web data are normally highly sparse in nature and formulating a personalized user profile was a difficult task. The previous methods used user based CF method to address the former issue while Extended Jaccard Similarity (EJS) was introduced to address the latter. Although these methods found to be effective, there were scopes for improvement in the Query processing time and accuracy. To solve this problem the proposed system design an Enhanced DBSCAN and TSVM based hybrid recommendation scheme. In the proposed system, the user interests and profile are analyzed by mining the web logs. After de-noising, user profile clustering is done through improved DB scan where Anarchic Society Optimization is used for parameter fine tuning. The sequential pattern of the users is evolved based on the GSP approach. The user interests are then evaluated using the TSVM which makes use of the weighted mean for the calculation of trust value of the users. Lastly, the personalized recommendations are generated based on the clicks, selections and browsing patterns. The proposed method is applied to MOVIE LENS data and found to perform better than the previous methods and
Introduction

In the digital era, with the rise in online web based services, recommender systems play a crucial role in day to day life. Recommender systems touch every aspect of online world starting from e-commerce to online advertisements which targets the interest of specific type of users. In generic manner, the recommender systems are the methodologies or the algorithms which aims in suggesting the most relevant items or services to a particular user (ex ; the movies to watch, the books to read, to choose a product, etc)[1-2]. The recommender systems have evolved as a critical parameter in many industries as it stands as vital in business expansion and acquisition of customers [3-4].

Traditionally, the recommender systems are categorized into collaborative and content based filtering approaches[5-6]. The usage of Hybrid approach has become more popular among the researchers that combine the filtering approaches for increased performance. There are always many hindrances in combining these filter techniques[7-8]. Even the same type of techniques when tried to hybridized faces issues. Hybrid approaches are possible in many ways: by generating the recommendations through collaborative and content based filtering and then combining the results; By incorporating the features of content based filtering in collaborative filtering and vice versa; or by a unified approach which combines both the methods. Various research have proven that the hybrid approach out performs better than the individual approaches[9]. These hybrid approaches are also proven to solve major issues such as the data sparsity and user profile clustering. In short, the hybrid approach solves the bottle neck in knowledge engineering in various knowledge based systems [10-11].

The previous methods used user based CF method to address the former issue while Extended Jaccard Similarity (EJS) was introduced to address the latter. Although these methods found to be effective, there were scopes for improvement in the Query processing time and accuracy. To solve this problem the proposed system design an Enhanced DBSCAN and TSVM based hybrid recommendation scheme. In the proposed system, the user interests and profile are analyzed by mining the web logs. After de-noising, user profile clustering is done through improved DB scan where Anarchic Society Optimization is used for parameter fine tuning. The sequential pattern of the users is evolved based on the GSP approach. The user interests are then evaluated using the TSVM which makes use of the weighted mean for the calculation of trust value of the users. Lastly, the personalized recommendations are generated based on the clicks, selections and browsing patterns. The proposed method is applied to MOVIE LENS data and found to perform better than the previous methods and other state of the art recommendation models in terms of MAE, Speed and Accuracy.
Literature Survey

Wu-Yang et al (2019) proposed a methodology that combines Singular Value Decomposition (SVD) and Hybrid Collaborative Filtering for an effective and optimized solution for the data sparsity issue. The SVD is used for bringing in a reduction in user-view matrix that are obtained through web usage mining[12]. Both of the low ranked matrix are then employed for generating item based and user oriented predictions. An automated framework for building an automatic in a real-time scenario is proposed. The recommendation engine performs in online real time scenario. The experimental results through the MOVIE LENS data prove the efficiency of the proposed method.

Hussein et al (2019) proposed the novel Hybrid framework which builds complex and context aware solutions in online recommendation. The proposed method is based on the concept of dynamic conceptualization [13]. The proposed framework was also meant for integrating the existing user or product information from outside sources such as social networks, Movie lens, etc. The hybrid method combines the extra ordinary characters of state of the art methods and Collaborative filtering methods. The evaluation is based on three conceptual foundations. The results indicate that the proposed method was efficient in terms of accuracy by reducing the loss and Mean Square error value.

Shreya et al (2019) improved the quality of the Movie recommendation system through a Hybrid approach that combines both the CF models using SVM as a classifier. A genetic model is also proposed and the comparative result showed significant improvement in terms of quality, scalability and accuracy of the Movie recommendation [14].

G M Sarene et al (2020) investigated that the performance of a recommender system is more affected due to over and above specifications, data sparsity and scalability[15]. The proposed method exploits the individual measures in predictive relevance that are computed for every user for every instances of interest which ensures high precision and recall values. Experimental results show that the advantages which are introduced through this recommender system has potential suggestions.

J. Grivolla et al (2021) proposed a hybrid recommender system which integrates the demographic of users and the characters of the item set interactions[16]. The proposed system is then evaluated on the MOVIE Lens data which included gene information and the user data which operates to provide tailor made discount to the customers. The proposed method included additional items and more information of the users which has more information and impact on the quality of information specifically where less interaction between the user profiles is available.

Palit R et al (2020) provided basic recommender system and how to implementation using the content and collaborative filter methods [17]. The proposed model takes the matrix of user ratings of movies as the top inputs and recommendations are generated. The proposed method also discusses the hybrid or mixture of recommender models for effective recommendations.
**Proposed Methodology**

The Proposed methodology is shown in figure 3.1

![Proposed Methodology Diagram](image)

**Data preprocessing**

The proposed method has various patterns of user style and behavior during testing the browsing styles of users and taking out their server blogs. Normally, the log file might include various unnecessary behaviors in extremely little period of times. One more problem is, in one pattern it might includes simply an action.
These are between the noises with the purpose of will be removed from the beta version of sequential datasets. The noise finding and elimination steps are extremely owing to the power of modification it has.

Algorithm 1: Denoised based on one source de noising

1. Given user u, content m, trial t=1,2,3, and \( > 0 \)
2. \( \text{Rut} = \{r_1, ..., r_k\}(\text{ratings} \neq 0 \text{ for user} u \text{ in one trial}) \)
3. \( \text{Rut}' = \{r_1, ..., r_k\}(\text{ratings} \neq 0 \text{ for user} u \text{ in different trial}) \)
4. Being 0 the value of the non rated item
5. for \( r_i \in \text{Rut} \) do
6. if \( 0 < |\text{Rut}(m) - \text{Rut}'(m)| > \gamma \) then
7. \( \text{Rut}(m) = 0 \) (SD: we delete rating)
8. else if \( 0 < |\text{Rut}(m) - \text{Rut}'(m)| > \gamma \) (MD: ratings in both trials differ by less than a given threshold) then
9. \( \text{Rut}(m) = M(\text{Rut}(m), \text{Rut}'(m)) \)
10. else
11. \( \text{Rut}(m) = \text{Rut}'(m) \)
12. end if
13. end for

The proposed work, given the case of varying thresholds, the Strong Condition of Disagreement(SD) is neglected from the current dataset.

Clustering Module

Recommendations shouldn't be made designed for the entire pool of users, since still for users by means of related specific interests; their ability toward handle a task is able to differ appropriate toward variations in their information level. The first step is taken using the clustering method and carried out as an step in achieving the user group based on the style of browsing they do. Depending on the various patterns of browsing style and users’ behaviors the clustering is carryout via the use of DBSCAN clustering algorithm.

DBSCAN is used to group the similar kind of user profiles. The DBSCAN method usually observes the user groups as dense regions of objects in the information space with the purpose are divided by means of regions of low-density objects. It is generally has two parameters such as Eps and MinPts as inputs for grouping similar Users. From the following description, it simply requires to find out each and every one the maximal density connected spaces to cluster the users in a search space.

Calculating the Parameters \( \epsilon \) and Min Pts
These steps which are dynamic provide couple of input in data pints which are auto created parameters which differ in the direction of data points that belong to the user.

The distance among two Users’ data points

The distance \( d(a, b) \) function measures the similarity of two Users’ data points.
d(a, b) = q√\left(μ_1 |X_{a1} - X_{b1}|^q + μ_2 |X_{a2} - X_{b2}|^q + ... + μ_p |X_{ap} - X_{bp}|^q\right) \quad (1)

Where, a = (x_{a1}, x_{a2}, \ldots, x_{ap}) and b = (x_{b1}, x_{b2}, \ldots, x_{bp}) are two p-dimensional Users’ data points, and q is a positive integer.

Automatic parameters creation dynamic method

The proposed DBSCAN algorithm discovers groups inside varied densities, by creating multiple pairs of ε and MinPts routinely.

Algorithm 1

Let X = \{x_1, x_2, x_3, \ldots, x_n\} is set of Users’ data points (users). DBSCAN requires three parameters: ε (eps), similarity and the smallest number of Users’ data points (users) are necessary to form a cluster (minPts).

Step 1: start with an arbitrary starting Users data points with the purpose of has not been visited

Step 2: Extract the neighborhood of this learner’s data points using ε.

Step 3: If there is similar browsing style of user around this Users data points (user) then clustering process starts and Users data points is marked as visited

Step 4: else Users data points is labeled as noise

Step 5: step 2 is repeated if all the data points are found to be a bit of cluster. This is done because of the anticipation that every data point inside a group are focused.

Step 6: A new visited Users’ data points (user) is retrieved and processed, leading in the direction of the discovery of a further cluster.

Step 7: This process continues until each and every one Users data points (users) are marked as visited.

The important parameters are ε and minPts which needs to be resolved by using Anarchic Society Optimization (ASO) algorithm[18]. The clustering accuracy of every part is calculated depending on the two parameters ε and minPts. As per the fitness which is computed, \(X^*(k), r_{i}^{best}\) and \(G_{best}\), the movement arrangement and a new value of the part will be found.

Algorithm 2: ASO Algorithm

Input: N parts , Output: Optimal parameters ε and minPts.
Users with diverse browsing styles have distinct rehashed succession arrangements. As a result, students were divided into groups based on their learning methods, and then personal behaviour standards were determined for each student using the Generalized Sequential Pattern algorithm (GPS) [19]. For comprehending continuous example mining concerns, a huge GPS is offered. One way to make use of the level-wise model is to first discover all of the continuous pupils in a level-wise approach. This effectively accounts for the verification of every singleton part’s events in the database. Non-frequent products are ejected in this way, and the transactions are channeled. These have couple of advances.

**Candidate Creation**

Given the set of frequency, and the frequencies, \( F_{n(i-1)} \), the candidates that are used for the following pass are then created using the combination of \( F_n \) by itself. Hen, the sub sequences which are not frequent are removed from the pruning.
**Pruning phase**

A pruning step removes some sequence at least one of whose series is not frequent.

**Support Counting**

Frequently, a hash tree–based search is proposed for capable support counting. Subsequently last non maximal frequent sequences are removed.

**Algorithm**

1. \( F_{ni} \) = the set of frequent l-sequence
2. \( i=2 \),
3. do while \( F_{n(i-1)} \)=Null;
4. Generate candidates sets \( C_{ni} \) (set of candidate l sequences);
5. For all j/p sequences Q in the database D do
6. Raise count of all a in \( C_{ni} \) if Q supports a \( F_{n}=\{a \in C_{ni} \text{ such that its frequency exceeds the threshold} \} i=i+1;
7. Result =Set of all recurrent sequences is the union of all \( F_{n} \)is
8. End do
9. END

**TSVM**

To know the Users’ data, the transductive support vector machine is proposed in this work. The Users’ rating may be described based on the % ideal answer, that are of dividing the clustered user based on the ratings of these frequent sequences[20]. TSVM is a continuous classifier that uses a transductive approach to locate the optimum hyper plane in the Users’ space The browsing step of the TSVM is able to be considered as issue of optimization and described as follows:

\[
J(w, \xi, \xi^*) = \frac{1}{2} ||w||^2 + C \sum_{i=1}^{l} \xi_i + C^* \sum_{j=1}^{d} \xi^*_j \tag{2}
\]

Will be focused to

\[
y_i((\varphi X_i).w+b) \geq 1-\xi_i, \xi_i \geq 0; i = 1,2, ..., l \tag{3}
\]

\[
y_j((\varphi X_j).w+b) \geq 1-\xi^*_j, \xi^*_j \geq 0; i = 1,2, ..., d \tag{3}
\]

Training the TSVM relates in the direction of handling the above mentioned issues. Lastly, the conclusion function of the TSVM with Lagrange multipliers \( \alpha_i \) and \( \alpha^*_j \) are described as follows,

\[
f(x) = \text{sgn} \left[ \sum_{i=1}^{l} y_i \alpha_i k(x,x_i) + \sum_{j=1}^{d} y^*_j \alpha^*_j k(x,x^*_j + b) \right] \tag{4}
\]

If two Users were calculated by the system by means of the parallel rating designed for a comparable navigational series, then they are supposed to be similar. Recommendation procedure is able to be performed depending on these
browsing styles which is performed based on the Personalized Browsing Approach (PLA).

**Recommendation process**

The student preferences and browsing behavior be able to be described by means of the subsequent three ways:

1. Clicks: This step is described as the small listing of material.
2. Selection: This step is described as material chosen and added toward cart.
3. Browsing: It is elaborated as below

All of these above steps are used for recognizing the Users comparative preferences $LP_{ij}$ used for every of the material referred from the information sources. The formula used for finding the $LP_{ij}$ is described as follows: 

$$LP_{ij} = \frac{LP_{ij}^c - \min_{1 \leq j \leq |M|}(LP_{ij}^c)}{\max_{1 \leq j \leq |M|}(LP_{ij}^c) - \min_{1 \leq j \leq |M|}(LP_{ij}^c)} + \frac{LP_{ij}^s - \min_{1 \leq j \leq |M|}(LP_{ij}^s)}{\max_{1 \leq j \leq |M|}(LP_{ij}^s) - \min_{1 \leq j \leq |M|}(LP_{ij}^s)}$$  

$LP_{ij}^c$, $LP_{ij}^s$ and $LP_{ij}^l$ describes the amount of references toward material during clicks, selection and browsing steps performed by the User 'i' in the material 'j' correspondingly. $\max_{1 \leq j \leq |M|}(LP_{ij}^c)$, $\max_{1 \leq j \leq |M|}(LP_{ij}^s)$, $\max_{1 \leq j \leq |M|}(LP_{ij}^l)$ describes the highest no. of clicks, selection and browsings for a performed by the User 'i' in the material ‘M’. $\min_{1 \leq j \leq |M|}(LP_{ij}^c)$, $\min_{1 \leq j \leq |M|}(LP_{ij}^s)$, $\min_{1 \leq j \leq |M|}(LP_{ij}^l)$ describes the lesser no. of clicks, selection and browsing performed by the User ‘i’ in the material ‘M’. 

The proposed personalized recommendation system is used toward improve user’s browsing knowledge.

**Experimental Results**

The experiment is deployed in Java with “oracle easy” as Database. The performance chart of the proposed Enhanced DBSCAN and TSVM based hybrid recommendation scheme is computed and distinguished with the existing state of the art methods and with the previous methods. The present and proposed methods are distinguished with respect to query processing time, Mean absolute error (MAE) metric, and accuracy. The MOVIE LENS Data is used for the experimental purpose which is open sourced and available from the data world. The proposed method DBSCN:TSVM is then compared with other state of the methods such as K-Means Neural Network (KNN), DBSCAN, Simple Support Vector Machine(SVM), Artificial Neural Network(ANN), Collaborative Filter Partial Timing (CFPT) and Extended Jacard Similarity (EJS)
The overall result analysis is shown in table 4.1

<table>
<thead>
<tr>
<th>Methods</th>
<th>Processing Time in Ms</th>
<th>Mean Absolute Error</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>0.92</td>
<td>0.96</td>
<td>0.69</td>
</tr>
<tr>
<td>DBSCAN</td>
<td>0.86</td>
<td>0.82</td>
<td>0.78</td>
</tr>
<tr>
<td>SVM</td>
<td>0.98</td>
<td>0.89</td>
<td>0.81</td>
</tr>
<tr>
<td>ANN</td>
<td>0.72</td>
<td>0.76</td>
<td>0.83</td>
</tr>
<tr>
<td>CPFT</td>
<td>0.68</td>
<td>0.62</td>
<td>0.79</td>
</tr>
<tr>
<td>EJS</td>
<td>0.86</td>
<td>0.73</td>
<td>0.80</td>
</tr>
<tr>
<td>DBSCAN:TSVM</td>
<td>0.41</td>
<td>0.25</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Table 4.1 Result analysis

### Query Processing Time

The query processing time is defined as the time taken for the recommendation set to get generated for a particular click, selection or search operation by the user. The experiment was conducted with 100 users as initial set which was slowly increased to 100000 users in a similar environment. The average time taken is calculated for each instance and is recorded. Figure 4.2 shows the comparative analysis of the query processing time of the proposed method with that of other state of the methods.
Figure 4.2 – Comparison of Query Processing Time

Figure 3 represents the comparison chart of query processing time performance for proposed Enhanced DBSCAN with TSVM based hybrid recommendation and existing methods. The proposed system accomplished better output than the current system, which is proved in the experimental result.

**Mean Absolute Error (MAE) metric**

The MAE is figured by summing these total blunders of the N which is proportional to ratings– forecast matches and afterward ascertaining the normal esteem. For each combine, \( <p_i, q_i> \) of anticipated appraisals \( p_i \) and the genuine evaluations \( q_i \), this metric joy the supreme blunder among them i.e., \(|p_i - q_i|\) similarly. To figure the measurable exactness, MAE metric is used, which is a calculation of the suggestions deviations from their actual User indicated values. Formally,

\[
MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N}
\]

(6)
Figure 4.3 represents the comparison output of the proposed enhanced DBSCAN and TSVM based hybrid recommendation and, existing DBSCAN with SVM based hybrid recommendation and trust based hybrid recommendation strategy with respect to MAE. The designed system accomplished better output than the current system, which is proved in the experimental result.

**Accuracy comparison**

The accuracy is calculated by initially calculating the True Positive(TP), True Negative (TN), False Positive(FP) and False Negative (FN) and through the calculation as shown below

\[
\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN} \quad (7)
\]

Figure 4.4 represents comparison of accuracy performance for proposed Enhanced DBSCAN with T SVM based hybrid recommendation and, existing methods. It is found that the proposed method outperforms the existing methods with an accuracy of 96%

![Comparison of Accuracy](image)

**Conclusion**

The recommender system framework is provided for a personalized item recommendation, specifically gives the recommendations to help the Users to identify and choose the effective information. In our work, cluster is done by Enhanced DBSCAN algorithm, which works according to the browsing style of the Users. Then T SVM algorithm computes the Users’ evaluation based on their habits and their interests. The student preferences and browsing behavior be able to be described by means of the subsequent three ways like clicks, selection and browsing. Finally, the experimental result shows that the designed system accomplishes good output when distinguished with the current system with respect to Processing time, MAE metric and accuracy.
References


