A Proposed λ_Mining Model for Hierarchical Multi-Level Predictive Recommendations

Yousef S. Alsahafi[1], Ayman E. Khedr[2], Amira M. Idrees[3]
University of Jeddah, College of Computing and Information Technology at Khulais,
Department of Information Technology, Jeddah, Saudi Arabia[1]
University of Jeddah, College of Computing and Information Technology at Khulais,
Department of Information Systems, Jeddah, Saudi Arabia[2]
Faculty of Computers and Information Technology, Future University in Egypt, Egypt[3]

***Abstract*—Delivering the most suitable products and services essentially relies on successfully exploring the potential relationship between customers and products. This immense need for intelligent exploration has led to the emergence of recommendation systems. In an environment where an immense variety exists, it is vital for buyers to own an intelligent exploratory map to guide them in finding their choices. Personalization has proven to be a successful contributor to recommenders. It provides an accurate guide to explore the users' preferences. In the field of recommendation systems, the performance of the systems has been continuously measured by their success in accurate, personalized recommendations. There is no argument that personalization is one key success; however, this research argues that recommendation systems are not only about personalization. Other success factors should be considered in targeting optimality. The current research explores the hierarchy map representing the strengths and dependencies of the recommendation systems pillars associated with their influence level and relationships. Moreover, the research proposes a novel predictive approach that applies a hybrid of content and collaborative filtering recommendation systems to provide the most suitable customer recommendations effectively. The model utilizes a proposed features selection approach to detect the most significant features and explore the most effective associations' schemes for the recommendations label feature. The proposed model is validated using a benchmark dataset by extracting direct and transitive associations and following the identified schematic for the required recommendations. The classification techniques are applied, proving the model's applicability with an accuracy ranging from 96% to 99%.**

***Keywords*—Recommendation systems; data mining; features selection; associations rules mining; classification techniques**

I. Introduction

Delivering the most suitable products and services essentially relies on successfully exploring the potential relationship between customers and products. This immense need for intelligent exploration has led to the emergence of recommendation systems [1]. The recommendation systems continuously apply different intelligent [2] as well as statistical [3] [4] techniques targeting accurate recommendations which directly lead to grasp the customer attraction and gain his positive interaction. The recommendation system types are content-based, knowledge-based, and collaborative filtering [5]. Each type has its own perspective in the analysis task to provide the most suitable recommendations. Content-based type focuses only on the product or service data, while collaborative filtering focuses on users' behavior. Moreover, knowledge-based recommendation systems provide recommendations based on the gathered knowledge of both of them, which could be considered the most challenging type as it should successfully force a homogeneous environment in which knowledge of both products and users could be mutually interpreted.

Although collaborative filtering is the most widely applied approach [6] [7], however, the current study highlights the competence of the knowledge-based approach over other types either the content-based or the collaborative filtering. Knowledge-based approach relies on both users' and products' data as well as these products' ratings. With the applicability to include this divergence of knowledge, the current study argues for higher and satisfied performance of the recommendation systems based on the knowledge-based approach. Although the recommendation systems field has had exceptional success, there are many challenges that should be considered. One of the challenges is to identify suitable methods for dealing with data sources and their issues, such as sparse data. Improving the scalability and efficiency of the recommendations is a continuous challenge in the field [8]. Many research papers have tackled these enhancements for one of the approaches; however, the integration of different approaches is not considered. The idea of collaboration has succeeded in different research to overcome one of the approach challenges with the other approach benefits. In addition, neighborhood techniques such as mining techniques, machine learning, and natural language processing contribute efficiently in the same direction. One of the neighborhood techniques is association rules mining. When these techniques are embedded in the recommendation systems, one of the substantial factors is the quality level of the explored associations. This quality level is heavily affected by the significance of the features [9]. The quality evaluation metrics, including support and confidence, may provide the same measures. However, this study argues that the recommendations could be strongly affected positively by the associations of the significant features. Therefore, one of the current research objectives is to explore the significant features association schemes with the label feature. In addition, considering the significant feature dependencies, even extend the associations' schemes targeting to provide more comprehensive and efficient recommendations. The associations' dependencies could be considered the pillar of building the transitive associations, contributing to more exploration for efficient recommendations.

In this study, the recommendations were based on a set of stages; each of these stages applies an intelligent technique and provides more exploration to the knowledge in-hand. The proposed recommender has three main stages. First, exploring the significant features through weighting methods with the collaboration of statistical methods. Second, exploring the direct association rules, then a second level of associations are explored for more accurate recommendations. Finally, build the semantic recommendation road map through the explored relationship between the features. The evaluation is performed by applying the classification algorithms to confirm the accuracy of the provided recommendations. In addition to the collaboration of the previously mentioned intelligent techniques, the contribution of the paper could be extended to the fact that the extracted map could be applied to different fields. The model is built on processing the provided data and is not restricted to a certain scheme. The remaining of the paper discusses the related work in Section II The proposed model details are discussed in Section III. The experimental study and the evaluation results are presented in Section IV. Finally, the discussion and the conclusion are presented in Section V and Section VI.

II. Related Work

Data mining and machine learning techniques have contributed to a wide range of developing intelligent recommendation systems over the years [10]. Many of the developed models achieved significant advancement in the field. These techniques can contribute from the start of the process [11] in the data pre-processing stage to the final stage of data presentation [12]. It has been proven by many research that the contribution of different models could provide more success to the recommendation process. One successful example is [13], whose aim was to recommend the price of the used cell phones based on the phone status using image processing techniques. Another research [14] followed the collaborative approach, which is also adopted in the current research, to recommend the targeted customers to grant bank loans. A survey was developed [15], which discussed the research for risk assessment and the customers' recommendation models. Another research in the banking sector [16], which reviewed the machine learning techniques that are utilized in detecting fraud in using credit cards. In the field of construction, an accurate budget recommendation was the objective of the research [17], which proposed a model for predicting the overhead cost for commercial projects. Most of the existing recommendation systems follow the seller-centric approach [18]. This approach is restricted to the seller's product availability, which could be more beneficial when extending this perspective to include a wider range of information to the user. While some research focuses on the products that are already bought by the user [19], however, extending the user focus with unintended yet related products could be a step forward in the recommendations.

A research in study [20] has presented recommendation systems using both data of the products and users' information. The proposed models utilized similarity measures for document analysis. However, the models suffered from the non-homogeneity of the items on focus during the analysis of the documents. There was also an issue in the model accuracy due to the non-dealing for the missing values. Another research in [21] depended on the users' response to the products' ratings. These models also applied similarity techniques as well as k-nearest neighbor mining algorithm. The model also suffered from the existence of synonyms while not manipulated. As the model depended on the user rating, the false rating was also considered a serious issue in the accuracy. Moreover, the research [22] focused only on demographic data such as age and gender. The model applied group recommendations using clustering techniques. However, the complete dependence on the demographic data is a major issue that hinders the accuracy. The research in study [23] proposed a content-based model for recommendations, which also depended on the user input. Although the model also highlighted the limited context input with a recommendation to the immense need for including other data perspectives and working with various types of data sources.

The research in study [24] and [25] highlighted the issue of including new items in the data sources whose ratings are not specified. This issue is called the cold start issue. The current research proposed the approach of the collaborative environment between both user and item data with pre-processing mining steps to solve this issue. Additionally, the issue of sparsity is also highlighted in the model proposed in the research [26] [27]. The prior research recommended a clustering approach to resolve this issue. While it could provide more accurate recommendations, however, the current research follows the approach of multi-leveling to further reach higher level of accuracy. The issues of scalability are highlighted in the research [28] and [29], which may be a result of the lack of user trustiness [30] and [31]. Both issues could be viewed as a dilemma situation. Therefore, working on the system performance by leveling up the users' confidence in the system recommendations is one of the motivations in the current research.

III. Proposed Model

The main target of the proposed model is to build the recommendation road map for the business on focus. The model presents the user characteristics, the possible multi-level recommendations for the users on focus as well as an exploratory map for potential users to enlarge the users' segment. The proposed model has four main stages. The following subsections discuss each stage and the included steps in detail.

A. Data Preparation Stage

1) Gather data: Data sources vary. It could be gathered from databases, users' responses, or real-time data from devices that measure and register data in a timely manner. Data could be static or streaming. These different sources of data are processed in different methods due to the difference in nature for these resources. The primary goal of this step "gather data" is the ability to collect the available data that could efficiently contribute to the recommendation main task. Data could be categorized as primary or secondary. The definition of both types varies according to the surrounding environment. One definition is the fact that the primary data is collected directly from the source itself such as the measuring devices while the secondary data is previously gathered for other reasons and could be utilized in the current task. Moreover, secondary data could be previously processed before delivery. In this section, general definition will be presented for both types as there is no restriction in the proposed model on either the data type or category. However, in the experiment section, the types of data will be explicitly highlighted.

Gathering data starts with identifying the potential data sources. The set of data sources (DS) is defined as follows:

$DS = \{ds1, ds2, \ldots.dsj | j \in \mathbb{N} \}$

DS represents the set of data sources that includes all the potential data sources. These sources could be primary or secondary. The sets of primary data sources (PDS) and secondary data sources (SDS) could be defined as follows:

$PDS = \{pds1, pds2, \ldots.pdsi | i \in \mathbb{N}, pdsi \in DS \}$

$SDS = \{sds1, sds2, \ldots.sdsk | k \in \mathbb{N}, sdsk \in DS \}$

$DS = PDS \cup SDS$

The general definition of the set of features for the data source dsk is as follows:

$Features(dsk) = \{fk1, fk2, \ldots. fks \mid k, s \in \mathbb{N}, dsk \in DS\}$

The general definition of the data sources with defining their types is as follows:

$TDS = \{<ds1, typet1>, <ds2, typet2>, \ldots.dsj | j \in \mathbb{N} \}$

2) Integrate multiple resources: The goal of data integration is enriching the model with the data from different data sources. Integrating data leads to a uniform data sources' access. Although the same feature could be a part of different data sources description, however, it could have different types. For example, the gender feature could be nominal in one source (male and female) and integer in another sources (1 and 2). Therefore, the unification process should consider a unifying pre-step for the features type. The following algorithmic steps should be applied.

```
h=1
Repeat
  For each dsh+1 in DS
        For each feature fhs in Features(dsk)
        If  fhs∈ fhs+1  & type (fhs) <> type (fhs+1)
                      unify features' types (fhs, fhs+1)
until h = k-1
```

Then, integrating the data sources requires unifying the descriptive features to allow a single view for all data. The proposed model focuses on transactional data sources. Therefore, the set of the unified criteria for all data sources after integration is as follows:

$Features(DS) = \{fks, fyt, \ldots. fwh \mid k,y,s,t,w,h \in \mathbb{N} \}$

$Features(DS) = Features(ds1) \cap Features(ds2)\ldots.. \cap Features(dsk)$

$Features(DS) = \cap_{l=1}^{k} Features\ (ds_l)$

3) Data cleaning: Also called "data wrangling", one of the vital steps in the data analytics process is data cleaning. The hygiene of data heavily affects the whole analytics process. The inconsistencies and existence of errors directly affect results and cause flawed data which could not be configured and leads to unreliable results. The following cleaning steps are performed.

Remove direct redundancies: two records are directly redundant when they have the same values for all of their attributes.

Consequently, the dataset dsk` represents dsk after removing the redundancies. dsk` could be defined as follows:

$dsk` = \{tn,\ldots tm\} | dsk` \subseteq dsk \mid \exists\ tg, tg \in dsk, tg \notin dsk`, tg = tt, tt \in dsk`$

Remove outlier: a record is defined to be a direct outlier if the values of its attributes do not match the standard values. It could be due to incorrect calculation or incorrect data entry. For example, a student record is an outlier if its id does not follow the faculty standards such as to be 4 digits instead of 6 digits and does not have the year as required.

The set of outliers' records of the set dsk could be identified as follows:

$Outlier(dsk`) = \{tn,\ldots tm\} | Outlier(dsk`) \subseteq dsk`$

Consequently, the dataset dsk`` represents dsk` after removing the outliers. It could be defined as follows

$dsk`` = dsk` \cap Outlier(dsk`)$

Remove contradictions: two records are contradictory if they provide conflicting data. For example, one record may include the birth date while the age does not match with this date, or the salary of the employee is less than the required salary taxes.

The set of contradictions' records of the set dsk could be identified as follows:

$Contradiction(dsk``) = \{tn,\ldots tm\} | Contradiction\ (dsk``) \subseteq dsk``$

Consequently, the dataset dsk``` represents dsk`` after removing the Contradictions. It could be defined as follows

$dsk``` = dsk`` \cap Contradiction(dsk``)$

Adapt missing values: different methods could be applied for missing values detection. A simple solution is to remove the record with the missing value while a more informative solution is to predict this value either with statistical calculation or with intelligent techniques.

B. *Explore the Features' Significant Level*

The goal of this stage is exploring the most informative features in the dataset. Different methods could contribute to this stage. Filtering, wrapper, and embedded categories are introduced in different research [32] [33]. The proposed model does not restrict the contribution to a certain category or method, rather, the selection of the contributing methods is based on the applying experiment. In this research, the contributing methods will be stated in the experiment section.

1) Set weighting measures: The set of the types of weighting methods is defined as follows:

WT = {T1, T2, T3}

The set of the weighting methods that belong to a one of these types is defined as follows

WM (T1) = {wT11, wT12…, wT1x}

WM (T2) = {wT21, wT22…, wT2y}

WM (T3) = {wT31, wT32…, wT3z}

The set of all weighting methods are determined which is defined:

W = {w1, w2, ….wz| z ∊ ℕ, wz ∊ [WM (T1) | WM (T2) | WM (T3)]}|

W = WM (T1) ∪ WM (T1) ∪ WM (T1)

The weighting threshold is identified to be equal 50% (0.5)

2) Apply weighting techniques: Each weighting measure is applied on the dataset and the features weightings are determined. The set of weighted features for each weighting technique is defined as

Wgt (wz) = {<fwh, wgt_valwhz>| wz ∊W, fwh∊ Features (DS), wgt_valwhz>=0, wgt_valwhz <=1}

Reversibly, each attribute is weighted by a set of weighting measures, the set of all weights for a defined attribute is as follows:

Fgt (fwh) = {<wz, wgt_valwhz>| wz ∊W, fwh∊ Features (DS), wgt_valwhz>=0, wgt_valwhz <=1}

The matrix of all features' weights with respect to the weighting measures is as follows:

FT-WT = $\bigcup_{k=1}^{j} \bigcup_{l=1}^{i} \mathrm{AttWt}\,(\mathrm{j}, \mathrm{i})$ where j is the feature index and i is the weighting measure index

3) Explore features' significant level: Exploring the significance of the datasets features is performed through the following steps. The exploration task is the pillar to identify the suggested associations' schemes which relies on the features according to their strength in affecting the prediction task performance. This section clarifies the steps for exploring the significance level of the features while the following sections continue to discuss how the prediction process is performed.

The first step is building the matrix for each feature with respect to the label feature y (FT-WT). This step results in identifying x-1 matrices where x is the number of features.

FT-WT (x) = $\begin{bmatrix} \mathrm{ft}_{z1} & \cdots & \mathrm{ft}_{y1} \\ \vdots & \ddots & \vdots \\ \mathrm{ft}_{zj} & \cdots & \mathrm{ft}_{yj} \end{bmatrix}$ where z ≠ y, z, y ≠ j

The matrix of each of the contributing weighting measures h highlights the contribution of a feature x on the consistency of feature y. This impact is determined by the weight of x that is determined by h given that y is the label attribute. The matrix of the weighting measure h is constructed as follows:

After applying weighting technique h on feature x with label feature y, the weights ranges are normalized by setting the value of 1 for the weights of range from 0 to <0.3, 2 for weights of 0.3 to <0.6, 3 for the weights of range 0.6 to 1, and 4 otherwise. The matrix wtxy element is the determined normalized value while the value of wtyx equals to the multiplicative inverse of the normalized value 1/ wtxy. The following matrix represents the normalized matrix of the weighting measure i for all attributes weights with considering y as the label attribute.

$$\mathrm{WTy\ (h)} = \begin{bmatrix} \mathrm{Nwt}_{11} & 1/\mathrm{Nwt}_{12} & \cdots & \mathbf{1}/\mathrm{Nwt}_{1j} \\ & \vdots & & \\ \vdots\ \mathrm{Nwt}_{1x} & \mathrm{Nwt}_{2x} & \ddots & \vdots \\ \mathrm{Nwt}_{1j} & \mathrm{Nwt}_{2j} & \cdots & \mathrm{Nwt}_{jj} \end{bmatrix}$$

Following the same approach, a set of matrices are built for each attribute while considering the significance determinants (SD). The current research follows the same approach of [12]. The research adopted only three measures for λ with an acceptable accuracy percentage which confirmed the applicability of the proposed approach. However, the current research extends the contributing measures seeking targeting to raise the consistency accuracy level. The current research extends the significance determinants for each feature to include the minimum and maximum weighting value among all the weighting measures, the mean and median value of all weightings, the upper and lower inter quartile range, the mean of these determined upper and lower values of all the weightings, and the value of mean for λ. Finally, the features' level of consistency is determined according to the number of consistent determinants. In case two features have the same count of consistent determinants, then the weighting value is considered. The following rule applies for constructing the features consistency ordered set.

The set of SD for each feature ftx is as follows:

SD (ftx) = {sdt,…sdr}

The ordered set of features based on consistency level = {ftx,… fty} where (|SD (ftx)| >= |SD (fty)| & min(Fgt (ftx)) > min(Fgt (fty)))

C. Associations Exploration Stage

The core stage of the proposed model is the mining phase. Mining phase employs different methods with a variety of nature and analysis procedures targeting to recognize the non-trivial embedded patterns. Data mining is continuously contributing to the business field for many aspects. One of these aspects which is the objective of the current research is the customer. Data mining could apply collaborating techniques targeting for customers beneficiary such as intelligent advertising, recommendations, detect fraudulent, and others. In the proposed model, associations rules mining, which is one of the data mining techniques, is employed to detect the embedded schemes in data. These schemes will then be utilized to explore both direct and indirect level associations. Then, the explored associations are employed to support the following recommendation phase. The following subsections discuss the main steps of the mining stage.

1) Determine direct associations: In this step, associations rules mining algorithm is applied to the dataset in order to govern the customers' perceptions and willingness. Many algorithms are introduced for this task; therefore, the following definition is generic. However, a defined algorithm will be stated in the experimental study with a clear justification for the reason for this selection. At this step, there are no restrictions in the associations' generation. The following steps prune the generated rules to explore the first step in detecting the most vital associations to the assigned task which is based on the metrics' threshold.

∃ Assc = Ass (ftxi → ftwi) ∧ fxi , ftwi ∊ Features(DS) ∧ Support (Assc) >= SU_Threshold ∧ Confidence (Assc) >= Con_Threshold → Add_ Assc (ftwi, Pre-Accept (ci))

2) Identify direct associations schemes: Identifying the targeted schemes follows the explored significance level of the dataset features. The first step is prioritizing the influencing features with respect to the required label feature based on the influencing level. The level of influence is determined according to the consistency degree of the feature on focus with the label feature based on the previous λ measures. The second step is identifying the associations schemes to have the influencing features as premise and the label feature as conclusion. The final step is pruning the detected schemes with considering only the schemes that have the influencing features with consistency above the determined threshold. These steps are formally described as follows:

∀ ftx, ftx ∈ Features (DS), label (ftx) → ∃ fty, influence (fty, ftx)

∀ ftz, ftx, influence (fty, ftx), influence (ftz, ftx), Value (significance (fty)) > Value (significance (fty)) → ordered_significance (ftx) = {ftz, ftz,….. | fty ≺ ftz }

∀ ftz, ftx, ftz ∈ ordered_significance (ftx), significance (ftz) > consistency_Threshold →

Schemes (ftx) = Schemes (ftx) ∪ association (ftz → ftx)

3) Filter direct associations: Based on the assigned schemes, this step focuses only on a subset of the associations that satisfy the required schemes as well as the minimum threshold of the performance metrics, support and confidence. The filtering step is described as follows:

Target = { ftx,,…., fty} where ftx ∊ Features (DS), fty ∊ Features (DS)

∀ ftx, ftx, ∈ Target, ASS = association (ftz → ftx), Support (association (ftz → ftx)) >= Min_Support, Confidence (association (ftz → ftx)) >= Min_Confidence →

4) Identify indirect associations schemes Based on significant level: In this step, an argument is highlighted for the ability to gain higher recommendation accuracy by considering transitive associations. A feature may be indirectly significant to the label feature is the following rule applies.

∀ ftx, fty, ftx, ∈ Target, ASS = association (ftz → ftx), Support (association (ftz → ftx)) >= Min_Support, Confidence (association (ftz→ ftx)) >= Min_Confidence , ASS = association (fty→ ftz), Support (association (fty→ ftz)) >= Min_Support, Confidence (association (fty→ ftz)) >= Min_Confidence → Indirect_ASS(ftx) = Indirect_ASS(ftx) ∪ ASS (fty→ ftz → ftx)

Therefore, extending the previous identifying the associations schemes step is performed in the current step. After the influencing features are identified, then they are considered as label features and the same previous steps are performed. Each influencing feature is considered a label, then schemes are determined, and associations are detected with the same previous description and pruning rules. The idea of this step is extending the opportunities for higher performance recommendations based on the existence of indirect influencing features. A feature x has an influence on feature y when there exist a third feature z that is influenced by x and, at the same time, influencing y. although the origin definition of the indirect relation states that this influence exists only through the feature z, however, this relation restriction is not considered in the current relation definition. This is because the main objective of extending the relationship is to consider the related factors for highlighting more opportunities in exploring different opportunities for the most available as well as applicable recommendations with highest accuracy.

5) Extract direct & indirect associations: The final step in the mining stage is extracting the relevant associations to the assigned task. based on the task, the association scheme is determined, and its corresponding associations rules are explored. These rules are then utilized in the recommendation stage.

D. Explore Recommendations Road Map

The deliverable of the proposed model is constructed in this stage. The road map of the recommendations is developed by applying a set of steps as follows.

1) Set recommendations threshold: One of the recommendations threshold (RTH) determination methods could be a direct method as it could be decided by the responsible. Other more intelligent methods could be applied such as the proposed method in [34]. It varies according to the applied domain. Some domains require a high level of accuracy such as healthcare or tourism. Other domains do not require accuracy to be critical such as restaurants or groceries. In this stage, according to the determined recommendations threshold, the associations are pruned while the recommendation threshold matches with the associations' minimum consistency level of the premises in the association rules. the following applied in this step

∀ ass in ASS, association (ftz → ftx), | consistency (ftz) > RTH → RecASS = RecASS ∪ association (ftz → ftx)

2) Detect recommendations peaks: The recommendation peak is defined as the point that has either the maximum or minimum features values with all other features with various values. Determining the recommendations peaks of a determined features ftx is performed by capturing the set of spike points of the feature x Spike (ftx). A value is identified as a spike point of a feature fx if it is associated with most of the values of the significant feature fy in the association rules set of fx and fy. Formally speaking, Spike (ftx) is constructed by the following rule.

∀ ftx, ∀ fty , ∀ Vali(fty), ∃ association (Vali (fty) → Vali (ftx)) ∈ ASS, → ∀ Vali (fty), feature(x,y) = feature(x,y) ∪ {<Count (association (Vali (fty) → Vali (ftx))), Vali (fty) , Vali (ftx)> }

Max_Count(association (Vali (fty) → Vali (ftx))) = Vali (ftx) where ∃ <Count (association (Vali (fty) → Vali (ftx))), Vali (fty) , Vali (ftx)> and ∀ Val(fty), ∀ j, j >=0, j < | ASS| , Count (association (Vali (fty) → Vali (ftx))), Vali (fty) , Vali (ftx)> > Count (association (Valj(fty) → Valj (ftx))), Valj (fty) , Valj (ftx)>

The spike of the feature ftx with respect to the significant feature fty is determined as: Spike(fty, ftx) = Vali (ftx)

The set of all spike points of the feature ftx is determined as: Spike (ftx) = $\bigcup_{y=0}^{|\text{Features(dsk)}|}$ Spike $(ft_y\,,ft_x)$

IV. Experimental Study

This section confirms the applicability of the proposed model. The section discusses the applied experiments and the evaluation results. The experiments dataset includes books data which aim is for books recommendations. The dataset describes the ratings of 271379 books from 278858 users with a total of 1048755 rating records [35]. The total number of features describing the dataset are fourteen features. As mentioned in the dataset source, the dataset was collected from Amazon and the books were purchased online. The dataset was distributed over three files. The first file describes the books data (ISBN, Title, Author, Publication Year, and Publisher). The total records were a total of 271289 records. The file included incomplete records, a set of these records were missing most of the features' values and the remaining set had only one of the features with a missing value. The first set of records was removed from the dataset, this set included 38876 incomplete records which were removed. The second set included 30341 records. These values were simply retrieved as most of the features' values exist. The second file represents the users' data which are described by the user id, location, district, country, and age. The file included the data of 276271 users. A set of 4563 records had incomplete residence data while a set of 109163 records had incomplete age data. This means that a set of 162545 records had complete data of users. According to this distribution, the experiment has been limited to 162545 users and 232413 books with a total rating records equal to 645212. After preparing the data in their original files, a simple code was developed to integrate the data into a unified structure.

The next step is identify the contributing weighting measures. Eight weighting measures are included in the experiments, they are Principle Component Analysis (PCA), Information Gain (IG), Information Gain Ratio (IGR), Support Vector Machine (SVM), and Correlation. The measures were applied and Table I presents the weighting results for the features above the threshold for each measure.

TABLE I. Weighting Measures Results for the Features above the Threshold

IG		IGR		Correlation	
rating	0.896	age	0.788	rating	0.959
rating	1.134	Pub. year	0.584	publisher	1.019
author	1.107	year	0.620	age	0.912
age	0.830	author	0.560	Pub. year	0.746
Pub. year	0.783	rating	0.812	country	0.798
				author	0.794
SVM		**PCA**			
country	0.762	publisher	0.959		
author	1.000	author	1.019		
age	0.889	age	0.912		
Pub. year	0.739	Pub. year	0.746		
		rating	0.798		
		country	0.794		

The next step is to detect the attributes' consistency. Each of the weighting technique contributed in all attributes' consistency detection. Five iterations are performed for each attribute with a total of seventy iterations. The percentage matrix is calculated by determining the product values and the Eigen vector value, then at least six of the λ determents determines should be consistent in order to consider the attribute to be consistent. Table II and Table III presents the Eigen vector and λ determents for the Author attribute respectively.

TABLE II. Eigen Vector for Author's Feature

	SVM	CRL	IGR	DEV	IG	P	EV
SVM	1.000	0.900	0.250	0.500	0.100	0.007	0.019
CRL	0.657	1.000	4.475	0.650	0.140	0.228	0.679
IGR	0.253	0.224	1.000	0.850	0.900	0.051	0.152
DEV	0.899	0.900	0.980	1.000	0.100	0.050	0.149
IG	0.100	0.146	0.655	0.106	1.000	0.001	0.700

TABLE III. Λ Statistics for Author's Feature

		L	C.I	C.R	C/NC
Lmax		5.9890	0.2473	0.2208	C
Lmean		4.8490	0.0378	0.0337	C
Lmedian		5.8240	0.2060	0.1839	C
Lmin		1.0550	0.9863	0.8806	NC
Range		4.6350	0.0913	0.0815	C
Standard Deviation		1.2990			
Inter-Quartile range	0.1770	1.2058	-1.0766	0.167	C
	0.4830	1.1293	1.0083	0.413	NC
Mean (rang,UpperQ)		5.2620	0.0655	0.0585	C
Mean (rang,LowerQ)		3.9420	0.2645	0.2362	NC
Mean (range, mean)		4.5620	0.1095	0.0978	C

The same calculations are performed for the eleven feature. Four attributes are considered consistent, they are (Author, Country, Age, Year, and rating). As a final step, with considering the weights of these features to the five weighting measures, it is detected that four of the consistent attributes are significant, they are (Author, Country, Age, and rating). Therefore, these are the attributes that will contribute to the associations' exploration stage. The next phase is identifying the required associations' schemes. Following the selected features, the associations' schemes are as follows: Age → Author, Author → Age, Country → Author, Author → Country, and Author → rating. As the main label feature is the rating, therefore, the indirect associations' schemes are detected to be Age→ Authors → rating, Country → Author → rating. The final step is to identify the recommendations map. The explored map has been discussed with professors in the literature field who confirmed the strong associations and dependencies. A second evaluation step is applying a set of classification techniques over the dataset with considering only the features that are contributing in the association. The first experiment is applying five classification techniques namely K nearest Neighbor (KNN), Naïve Bayes (NB), Enhanced ID3 (EID3) and the contributing features are the age and authors with the rating as the label feature. The second experiment includes the country and authors with also the rating as the label feature. Moreover, a third experiment is conducted that includes the three features. The results reveals the high accuracy percentage which confirms the applicability of the explored associations and consequently the proposed model. Table IV demonstrates the average evaluation results for the classification experiments.

TABLE IV. CLASSIFICATION RESULTS

Criteria	KNN	NB	EID3
Accuracy	95.96 %	97.98 %	98.99
Classification Error	4.04 %	2.02 %	1.01
Kappa Statistic	0.937	0.969	0.984
Weighted mean recall	96.17 %	97.94 %	99.28
Weighted mean precision	95.56 %	97.52 %	98.85
correlation	0.928	0.956	0.972

V. DISCUSSION

According to the results, the proposed approach succeeds to perform effective feature selection with a generality to be applied in different fields. The presented performance measures revealed the success of the approach, however, the missing percentage in the classification task could be due to the limited number of attributes in the dataset. Datasets with higher number of attributes could results to higher performance.

VI. CONCLUSION

This research proposed an intelligent model that aims to build the recommendation road map for business. The model explores the possible multi-level recommendations for the users on focus as well as an exploratory map for potential users to enlarge the users' segment. The proposed model had four main stages which applies mining techniques in the preprocessing and explorations stages, features selection techniques for exploring the significant features and statistical techniques to more illumination for these features in order to provide the most accurate recommendations base on these features associations. The proposed model has been evaluated by applying the detailed stage on a dataset of books. The results have been evaluated by experts and by applying classification models and evaluate the classification results. The results revealed a success in the classification that ranged from 96% to 99% which ensures the model applicability. More future work could be further applied to the proposed model. Another future direction to conduct more experiments could be conducted from different fields. Investigating more methods for feature selection could be further developed. Another enhancement direction is including the users' comments and apply text mining techniques with embedding the keywords as features.

ACKNOWLEDGMENT

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under grant No. (UJ-23-DR-13). Therefore, the authors thank the University of Jeddah for its technical and financial support.

REFERENCES

[1] A. Al Mazroi, A. E. Khedr and A. M. Idrees, "A Proposed Customer Relationship Framework based on Information Retrieval for Effective Firms' Competitiveness," Expert Systems With Applications, vol. 176, 2021.

[2] Y. Helmy, A. E. Khedr, S. Kholief and E. Haggag, "An Enhanced Business Intelligence Approach for Increasing Customer Satisfaction Using Mining Techniques," International Journal of Computer Science and Information Security, vol. 17, no. 4, 2021.

[3] A. M. Idrees and W. H. Gomaa, "A Proposed Method for Minimizing Mining Tasks' Data Dimensionality," International Journal of Intelligent Engineering and Systems, vol. 13, no. 2, 2020.

[4] M. Atia, M. Mahmoud, M. Farghally and A. M. Idrees, "A Statistical-Mining Techniques' Collaboration for Minimizing Dimensionality in Ovarian Cancer Data," Future Computing and Informatics Journal, vol. 6, no. 2, 2021.

[5] A. M. Idrees, A. E. Khedr and A. A. Almazroi, "Utilizing Data Mining Techniques for Attributes' Intra-Relationship Detection in a Higher Collaborative Environment," International Journal of Human-Computer Interaction, 2022.

[6] A. M. Idrees and F. K. Alsherif, "A Collaborative Evaluation Metrics Approach for Classification Algorithms," Journal of Southwest Jiaotong University, vol. 55, no. 1, pp. 1-14, 2020.

[7] F. Abogabal, S. M. Ouf and A. M. Idrees, "Proposed framework for applying data mining techniques to detect key performance indicators for food deterioration," Future Computing and informatics journal, vol. 7, no. 2, 2022.

[8] M. Tamer, A. E. Khedr and S. Kholief, "A Proposed Framework for Reducing Electricity Consumption in Smart Homes using Big Data Analytics," Journal of Computer Science, vol. 15, no. 4, 2019.

[9] M. Attia, M. A. Abdel-Fattah and A. E. Khedr, "A proposed multi criteria indexing and ranking model for documents and web pages on large scale data," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 10, 2022.

[10] A. E. Khedr, A. I. El Seddawy and A. M. Idrees, "Performance Tuning of K-Mean Clustering Algorithm a Step towards Efficient DSS," International Journal of Innovative Research in Computer Science & Technology (IJIRCST), vol. 2, no. 6, pp. 111-118, 2014.

[11] D. H. A. Hassouna, A. E. Khedr, A. M. Idrees and A. I. ElSeddawy, "Intelligent Personalized System for Enhancing the Quality of Learning," Journal of Theoretical and Applied Information Technology, vol. 98, no. 13, pp. 2199-2213, 2020.

[12] A. M. Idrees, A. I. ElSeddawy and M. O. Zeidan, "Knowledge Discovery based Framework for Enhancing the House of Quality," International

Journal of Advanced Computer Science and Applications (IJACSA), vol. 10, no. 7, pp. 324-331, 2019.

[13] A. M. Idrees and S. Taie, "Online Price Recommendation System for Shopping Used Cell Phones," Research Journal of Applied Sciences, Engineering and Technology, vol. 13, no. 1, pp. 15-23, 2016.

[14] A. M. Idrees and A. E. Khedr, "A Collaborative Mining-Based Decision Support Model for Granting Personal Loans in the Banking Sector," International Journal of E-Services and Mobile Applications (IJESMA), vol. 14, no. 1, pp. 1-23, 2022.

[15] M. sharaf, S. Ouf and A. M. Idrees, "Risk Assessment Approaches in Banking Sector-A Survey," Future Computing and Informatics Journal, vol. 8, no. 1, 2023.

[16] N. S. Elhusseny, S. M. Ouf and A. M. Idrees, "Credit Card Fraud Detection Using Machine Learning Techniques," Future Computing and Informatics Journal, vol. 7, no. 1, 2022.

[17] A. H. Z. Hassan, A. M. Idrees and A. I. Elseddawy, "Neural Network-Based Prediction Model for Sites' Overhead in Commercial Projects," International Journal of e-Collaboration, vol. 19, no. 1, 2023.

[18] S. Fazeli, H. Drachsler, M. Bitter-Rijpkema, F. Brouns, W. van der Vegt and P. B. Sloep, "User-centric evaluation of recommender systems in social learning platforms: accuracy is just the tip of the iceberg," IEEE Transactions on Learning Technologies, vol. 11, no. 3, 2018.

[19] A. M. Idrees and E. Shaaban, "Reforming home energy consumption behavior based on mining techniques a collaborative home appliances approach," Kuwait Journal of Science, vol. 47, no. 4, 2020.

[20] Z. Yang, B. Wu, K. Zheng, X. Wang and L. Lei, "A survey of collaborative filtering-based recommender systems for mobile internet applications," IEEE Access, vol. 4, 2016.

[21] C. He, D. Parra and K. Verbert, "Interactive recommender systems: a survey of the state of the art and future research challenges and opportunities," Expert Systems with Applications, vol. 56, 2016.

[22] M. Elahi, F. Ricci and N. Rubens, "A survey of active learning in collaborative filtering recommender systems," Computer Science Review, vol. 20, no. C, 2016.

[23] N. Pereira and S. L. Varma, "Financial planning recommendation system using content-based collaborative and demographic filtering," Smart Innovations in Communication and Computational Sciences, Part of the Advances in Intelligent Systems and Computing, vol. 669, 2018.

[24] V. R. Revathy and A. S. Pillai, "A proposed architecture for cold start recommender by clustering contextual data and social network data," Computing, Communication and Signal Processing, Part of the Advances in Intelligent Systems and Computing, vol. 810, 2018.

[25] Y. Zhu, J. Lin, S. He, B. Wang, Z. Guan, H. Liu and D. Cai, "Addressing the item cold-start problem by attribute-driven active learning," IEEE Transactions on Knowledge and Data Engineering, vol. 32, 2020.

[26] S. Ahmadian, M. Afsharchi and M. Meghdadi, "A novel approach based on multi-view reliability measures to alleviate data sparsity in recommender systems," Multimedia Tools and Applications, vol. 78, 2019.

[27] A. K. Sahu and P. Dwivedi, "User profile as a bridge in cross-domain recommender systems for sparsity reduction," Applied Intelligence, vol. 49, no. 7, 2019.

[28] H. Varudkar, S. M. Deosthale and J. Mehta, "Collaborative recommendation system based on Hadoop," Global Journal for Research Analysis, vol. 6, no. 1, 2016.

[29] V. Koshti, N. Abhilash, K. S. Gill, N. Nair, M. B. Christian and P. Gupta, "Online partitioning of large graphs for improving scalability in recommender systems," Computational Intelligence: Theories, Applications and Future Directions, vol. 2, 2019.

[30] H. Feng and T. Tran, "Context-aware approach for restaurant recommender systems," Encyclopedia of Information Science and Technology, Fourth Edition, 2019.

[31] M. Singh, H. Sahu and N. Sharma, "A personalized context-aware recommender system based on user-item preferences," Data Management, Analytics and Innovation, vol. 140, 2019.

[32] N. Hegazy, M. Khafagy and A. E. Khedr, "Proposed Approach for Academic Paper Ranking Based on Big Data and Graph Analytics," Journal of Theoretical and Applied Information Technology, vol. 101, no. 2, 2023.

[33] H. Elmasry, A. E. Khedr and H. M. Abdelkader, "Enhancing the Intrusion Detection Efficiency using a Partitioning-based Recursive Feature Elimination in Big Cloud Environment," International Journal of Advanced Computer Science and Applications,, vol. 14, no. 1, 2023.

[34] E. Hikmawati, N. U. Maulidevi and K. Surendro, "Minimum threshold determination method based on dataset characteristics in association rule mining," Journal of Big Data, vol. 8, no. 146, 2021.

[35] O. Monk, "https://www.kaggle.com/datasets/saurabhbagchi/books-dataset," 2004. [Online]. [Accessed 9 2023].