CD-SemMF: Cross-Domain Semantic Relatedness Based Matrix Factorization Model Enabled with Linked Open Data for User Cold Start Issue

Senthilselvan Natarajan1, Subramaniyaswamy Vairavasundaram1, Ketan Kotecha2, Indragandhi V3, Saravanan Palani1, Jatinderkumar R. Saini4, Logesh Ravi5

1School of Computing, SASTRA Deemed University, Thanjavur, India
2Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune, India
3School of Electrical Engineering, Vellore Institute of Technology, Vellore, India
4Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University), Pune, India
5School of Electronics Engineering (SENSE), Vellore Institute of Technology, Chennai, Tamilnadu, India

Corresponding author: Ketan Kotecha(head@scaai.siu.edu.in) & Subramaniyaswamy Vairavasundaram(svsubramaniyaswamy@gmail.com)

ABSTRACT Personalized recommendations to cold start user is one of the significant challenges in information filtering systems. Most of the existing systems inherited the idea of collaborative filtering (CF) and avoids the item metadata. For instance, consider a book domain whose metadata are author, publisher, language...etc. The elimination of item metadata is due to the diverse nature of item attributes in a cross-domain environment. Because of this ideology, it is hard for the filtering systems to provide better recommendations to cold start users. Cold start users are people who are new to a domain whose preferences are unknown. A novel model is proposed in this research work called Cross-Domain Semantic Relatedness based Matrix Factorization model (CD-SemMF) resolves the user cold start issue in collaborative filtering recommender system by exploiting Linked Open Data (LOD). “DBpedia” is a widely used LOD resource that contains semantic information about different domains, which is used in this research work to resolve the above-said problem. Here, the metadata available in LOD connects the items preferred by the target user from various domains. The fundamental knowledge graph links items from various domains and also benefits from cross-domain information. “LOD Semantic-Relatedness Measure” a new measure is proposed which calculates the closeness of items across domains. Semantic relatedness is measured instead of similarity in this research because the cross-domain item attributes are diverse. The Alternating Least Square method is applied here to learn the user preferences. The proposed model provides relevant, personalized recommendations for the target new user with the user preferences gained from the source domain and by exploiting item semantic relatedness. Experimental evaluation is done on Facebook and Amazon datasets. It is observed from the result that the proposed CD-SemMF model gives better recommendations in target domain for new users than the baseline methods.

INDEX TERMS Linked Open Data, Cross-Domain Recommender System, Semantic Relatedness, Item Metadata, and User Cold Start.

I. INTRODUCTION
Providing personalized items to the target user is one of the key challenges in the recommender system. With the emergence of the Cross-Domain Recommender System (CDRS), the major challenge in information filtering is reduced especially in user cold start problem [1]. In an aim to alleviate the data sparseness, cross-domain recommendation exploits information like item attributes and user preferences from distinct but related target domains. Almost all the existing cross-domain recommendation models are working based on the collaborative filtering (CF) algorithm [2]. CF approach exploits user preferences to connect target and source domains but ignores item contents. Consequently, they benefit from the fact that there is no necessity to analyze the contents of an item that are highly heterogeneous across domains and their inter-domain relationships are also hard to create. Semantic Web initiative [3] helps to address the heterogeneity problem. In specific, its reference implementation Linked Open Data [4] creates a huge number of publicly available inter-linked knowledge repositories on the Web, followed by the Semantic Web standards for data access and data representation. Presently, the web contains a broad array of structured data with item contents of different domains like science, history, media, arts, sports, industry, and so on. It not only contains the multimedia content and its related metadata, but also the semantic relationship existing between items and their metadata. Due to the accessibility of
item metadata and its semantic relationship from the LOD cloud, the major challenges of cross-domain recommendation are solved which are addressed here. This is achieved not only by focusing on item contents and user preferences but also by exploiting the content-based relationship between items of various domains.

The proposed model CD-SemMF uses the features of the LOD semantic web and its relationship as inter-domain links to transfer knowledge among different domains. It also enables cross-domain semantic relatedness between items which helps to generate relevant recommendations for new users in the target domain. Existing works like graph-based algorithm [5,6] addresses the recommendation challenge in heterogeneous datasets. They analyze the topology of the semantic network and jointly exploit both the item metadata and user preferences for effective recommendations. Unfortunately, they suffer from high computations due to increased semantic network size. But in the proposed model, the above-mentioned issue is resolved by introducing two novel steps. (1) Firstly, inter-domain semantic relatedness is computed from the semantic network which links items from various domains. (2) Secondly, the computed relatedness is used in the matrix factorization method to generate relevant recommendations. Hence, it is not necessary to deal with the complete semantic network for recommendation generation. Initially, each item in the dataset is mapped into entities of LOD to obtain the semantic metadata of various items in terms of SPARQL queries. Finally, attributes and their relations are extracted to improve the item profiles. A novel semantic relatedness metric is proposed called LOD semantic relatedness metric, which finds the inter-domain item relatedness. It is observed that the proposed model will be used effectively when the knowledge graph of fundamental LOD encodes direct/indirect connections existing between items of different domains. Relatively, the connections are general for knowledge domains like music, movies, and books with some amount of information overlap. Therefore, the key contribution of this research work is the development of LOD semantic relatedness metric with a novel and effective matrix factorization model “CD-SemMF” which exploits both item's metadata and user preferences for cross-domain recommendations in a cold start scenario. Further, Pilászy’s fast learning algorithm [7] is adopted in this work to build the model more effectively. The proposed model is evaluated on Facebook and Amazon datasets. Based on the target and source domain involved in the recommendation process, it is demonstrated that the proposed model generates more accurate recommendations than the baseline approaches in severe cold start scenarios.

Section II discusses the related work concerning cold start issues and cross domain recommender systems. Section III elaborates the working of the proposed Cross-Domain Semantic Relatedness Based Matrix Factorization Model for Cold Start User. Section IV illustrates the experimental evaluation of methods in the cold start scenario. Finally, Section V summarizes the contributions made in this research work and its future scope.

II. RELATED WORKS

"Cross-Domain" has recently become a possible solution to the problem of cold start users and data sparsity in the "recommender system". In diverse research fields, cross-domain recommendation issues have been discussed from different perspectives. The system is adjusted in a user-defined simulation as a feasible solution to reduce cold start problems [8-10] and bottlenecks in the recommendation system [11], as well as the practical application of knowledge transfer in machine learning [12]. Fernandez [1] focuses on how to use knowledge in cross-domain recommendation systems, and classifies the existing work as two-levels: one is a knowledge base, and the other is a knowledge transfer relationship, which is discussed below.

A. KNOWLEDGE AGGREGATION FOR CROSS-DOMAIN RECOMMENDER SYSTEM

Information from various domains are combined to provide a better recommendation for a new user in the target domain. Depending on the stage in the recommendation process where the aggregation is performed, it was distinguished into three situations. In the former case, these methods combine settings such as mark level, transaction log, and click data. The scoring matrix of multiple domains can be used to perform aggregation [13]. Generic representations are used for user preferences such as social tags or semantic concepts [14], which are linked to multiple domains through graphs [15]. User preferences are features that are not domain-related, such as personality features or assigned user elements [16]. In another scenario, user modeling information is transferred from numerous recommender systems to improve target recommendations. For example, several researchers import the user-defined neighborhood and user similarity calculated in the original area to the target domain [17]. Finally, some methods combine recommendations for various fields, by evaluating estimates and ranking probability distributions [18].

B. KNOWLEDGE TRANSFER FOR CROSS-DOMAIN RECOMMENDER SYSTEM

Establishing communication or knowledge transfer among domains supports recommendations. In this scenario, various technologies have been identified that use common sense to link domains, such as (1) element attributes, association rules, semantic networks, and cross-domain relevance [19]. (2) The method of sharing hidden attributes between the factor models of the source domain and the target domain can be achieved by using the same model parameters in the two factorizations [20-22] or by introducing new parameters of the extended factorization [23-25]. (3) A method of transmitting the evaluation template extracted by grouping the evaluation matrices of the source domain and applying it to the target domain. Once the problem is defined, Pan [26] proposed three unique collaborative recommendation information transfer strategies with additional data (TL-CRAD): (1) adaptive, (2) collective, and (3) integrated
knowledge. For this work, the authors studied papers related to different knowledge strategies, focusing on dissemination through predictive rules, regularization, and restrictions.

C. USER COLD START ISSUE IN RECOMMENDER SYSTEM

Most of the existing models focus on improving accuracy by reducing data dispersion [27,28]. In many areas, the average number of comments per user and per article is smaller, which may negatively affect the quality of recommendations. Data collected outside the target area can increase the density of classification which improves recommendations quality. Researchers seek to improve the user models that may have personalization-oriented advantages, such as (1) recognizing new user preferences for target domains [29], (2) improving the similarity between users and elements [30], and (3) Measuring vulnerabilities in social networks [31,32]. By importing settings from different sources outside the target domain, cross-domain techniques have also been applied to guide recommendation systems, and it has been suggested to improve the types of recommendations by providing a better range of user settings. Liu [33] proposed a neighborhood-based algorithm to resolve the user cold start issue but for a huge dataset, it struggles to give a good recommendation. A cluster-level algorithm was proposed which maintains high-quality relationships in cold start scenarios [34]. First, the author used skew matrix decomposition to map the evaluation matrix into a smaller hidden space. After this process, the k-means grouping algorithm is applied to classify users and items. A new way of customization between systems is designed which is based on two assumptions: there is a custom model that can be shared between platforms, and a particular system can support the custom model created by its system [35].

The recommendation model is based on matrix factorization enabled with latent semantic analysis (LSA), which is widely used in natural language processing and information retrieval. LSA attempts to derive implicit concepts automatically in text documents by using truncated singular values to approximate the term-document matrix. The first recommended matrix factorization method takes the same idea and applies it to the custom element matrix in the rating prediction problem [36]. In contrast to LSA, singular value decomposition (SVD) is poorly defined for sparse matrices commonly found in recommendation systems. Therefore, previous methods rely on classification techniques to fill in the missing matrix elements before applying SVD. The subsequent methods aim to consider only the observed levels and not the entire matrix because it leads to poor results. The model proposed by Funk is a popular method in this direction [37]. Hu [38] proposed an alternating least squares (ALS) method, which can effectively deal with missing values. An important observation of ALS is that when a set of parameters is set and other parameters are estimated, the optimization problem becomes convex and can be solved analytically using traditional least-square estimation. Hu introduced a MF model called iMF, which is expanded and improved in subsequent work. Takács [39] extends ALS to the rank-based matrix factorization method, which learns to predict the relative order of items, not individual levels. Paquet [40] proposed a graph-based Bayesian model, which can record the meaning of missing values so that the users who don't like articles can be distinguished from users who don't know the article. Using ALS and gradient descent methods, the decomposition model can be easily trained to reduce prediction errors. Once the problem is formulated as the main goal of the recommendation, they can be trained using the "Bayesian personalized ranking criterion" (BPR) [41]. Kabbur [42] proposed the FISM model, which is used to study the similarity matrix of the object as the product of two low-dimensional latent factor matrices. To solve the problem of the first N suggestions, the author uses the Bayesian ranking heuristic ranking loss, which optimizes the area under the curve. Many models represented so far can't provide a better recommendation to a new user. This is due to the ignorance of item metadata in the matrix factorization process and the usage of similarity measures in a heterogeneous environment.

III. CROSS-DOMAIN SEMANTIC RELATEDNESS BASED MATRIX FACTORIZATION MODEL (CD-SEMMF)

Conventional methodologies used graph-based techniques to exploit semantic metadata of items, but the proposed model initially computes the inter-domain content-based semantic relatedness between items. With this relatedness data, the proposed model leverages the parameters of the target and the source domains through matrix factorization. The proposed semantic relatedness-based matrix factorization model creates different assumptions about the target and source domains item latent factors’ relationship. Concurrently, it exploits both the item metadata and user preferences. Furthermore, Pilászy’s fast Alternating Least Square (ALS) training algorithm was adapted in MF to tackle the increased computational complexity of the proposed model, because it not only learns the user preferences for recommendations but also exploits source and target domains item metadata to connect various domains. Items of various domains are probably having varied attributes, so they are not directly associated with one another. For example, a book is described by the attributes author or book genres, while a movie is characterized by its director, movie genres, or cast. Content-based features are different for different domains, though they are conceptually associated with movie genres and book genres. Their features are not properly aligned, for example, humor books vs. funny movies. To address this weakness i.e. heterogeneity of item features from various domains, Linked Open Data is exploited in the proposed model to link entities from various domains. In specific, each item in the dataset is mapped to its corresponding DBpedia (LOD) entities. DBpedia is a multi-domain data storage, from there the structured and the semantic knowledge is acquired for processing. Mapping of item attributes to DBpedia semantic entities is described in section IV 1. After the
successful completion of mapping, DBpedia is used to calculate the semantic relatedness between entities. With that available semantic relationship, the structure of the graph and its attributes are mined. In specific, information from DBpedia is used to compute the semantic relatedness matrix between the target (IT) and source domain items (IS). Here SR stands for semantic relatedness and $SR \in R^{IT \times IS}$ where R is the user preference matrix. This proposed measure is called LOD semantic relatedness measure, its equation is shown below:

$$SR_{ij} = rtn(i, j), i \in I_S \text{and} j \in I_T$$  \hspace{1cm} (1)

In equation 1 ‘rtn’ is the relatedness calculation of each item’s attributes between source and target domains taken from LOD which is discussed in section III 1. Then the resultant inter-domain relatedness between items is used to relate different domains to generate cross-domain recommendations. When there are only a few rated items in the target domain for a particular user or in the cold-start scenarios, the proposed recommender model will suggest items that are semantically analogous to previously preferred items by the same user in the source domain. Thus, the model will be efficient only if there are some user overlaps among domains. This is because even the new users of the target domain may have rated a few items (user preferences) in the source domain. The working principle of the proposed recommender model is explained in detail in the subsequent sections which exploit semantic relatedness to regularize the factors of an item in matrix factorization. With the intention that related items tend to have similar parameters though they are from different domains [1]. For example, consider utensil and eatables domains where spoon and ice scream are taken for comparison. Even though it belongs to different domains some properties like serve, eat, and health are common among them. Even user preferences are unknown in the target domain; the proposed recommender model will suggest items to cold start users which are based on the preference in the source domain by the same user. The big picture of the proposed model is shown below in figure 1. The working description of this model is given in the subsequent sections III A, III B and III C.

![FIGURE 1. Architecture of proposed CD-SemMF Model](image-url)
A. LINKED OPEN DATA SEMANTIC RELATEDNESS MEASURE

Various applications like automatic text correction, information retrieval, word sense disambiguation, automatic indexing, and so on uses similarity measures or relatedness measures. However, these two terms are different from each other. Semantic relatedness is described as the degree to which two different words are related to each other by means of any of the semantic relationships like synonymy, hyponymy associative, functional, hypernymy, or meronymy. While the semantic similarity is the special kind of semantic relationships like synonymy, hyponymy associative, functional, hypernymy, or meronymy. The semantic similarity is the special kind of relatedness which considers only hyponymy or hypernymy relationships. On the other hand, semantic relatedness between any two words is not discovered directly but by the set of related words which are well associated and are named as the set of determinative words. Determinative words are closely related to those exact words. Inspired by Mehmet’s relatedness measure [43], with some improvisation a semantic relatedness measure is proposed which is incorporated in the proposed model to resolve the user cold start issue in CDRS. To calculate semantic relatedness, information like hierarchical relationships and other types of relationships between words are considered. Such information is collected from LOD and hence the model is termed as the LOD semantic relatedness measure. The relatedness measure uses the combination of word relations based on their importance or context. To demonstrate the variation among relatedness and similarity here presents the illustration of spoon and ice cream for semantic relatedness. Though they are not identical, they share some common features. In terms of functional perspective, both are closely related. For instance, people eating ice cream with a spoon. The proposed semantic relatedness measure working is described as follows.

Let A and B be two words, then relatedness between these two words are measured in the following ways:

\[ \text{rtn}(d_i, A) = \frac{\text{fre}(d_i, A)}{\max f_1} \]
\[ \text{rtn}(d_i, B) = \frac{\text{fre}(d_i, B)}{\max f_2} \]

Where \( \text{fre}(d_i, A) \) represents the total number of pages in web link where the determinative \( d_i \) and the word A occurs together. Similarly, \( \text{fre}(d_i, B) \) be the number of pages in web link where the determinative \( d_i \) and the word B occur together. Likewise, \( \max f_1 \) and \( \max f_2 \) are calculated as follows:

\[ \max f_1 = \max \text{imun} \{ \text{rtn}(d_1, A), \text{rtn}(d_2, A), \ldots, \text{rtn}(d_k, A) \} \]
\[ \max f_2 = \max \text{imun} \{ \text{rtn}(d_1, B), \text{rtn}(d_2, B), \ldots, \text{rtn}(d_k, B) \} \]

If the determinative \( d_i \) is highly associated with the word A or B, then the possibility of occurrence of that determinative on the page of that web link is relatively high. Particularly, if the determinative \( d_i \) is synonymy or closely synonymy to the word A or B, then \( \text{rtn}(d_i, A) = 1. \) Similarly, \( \text{rtn}(d_i, B) = 1. \)

STEP 3

Finally, the relatedness between words is calculated as follows:

\[ \text{rtn}(A, B) = \left( \sum_{i=1}^{k} \frac{\delta_i Q_i}{1+Q_i} \right) + s \left( \frac{1}{1+s} \right) \]

Here \( Q_i \) is estimated as follows

\[ Q_i = \frac{\min \text{imun} \{ \text{rtn}(d_i, A), \text{rtn}(d_i, B) \}}{\max \text{imun} \{ \text{rtn}(d_i, A), \text{rtn}(d_i, B) \}} \]

Then the co-occurrence factor \( \delta_i \) is described as follows

\[ \delta_i = \begin{cases} 2, & \text{if } d_i \text{ is occurred in both words A and B} \\ 1, & \text{otherwise} \end{cases} \]

Finally, ‘s’ denotes the synonymy factor which is described as follows:

\[ s = \begin{cases} 1, & \text{A and B are synonymy or nearly synonymy} \\ 0, & \text{otherwise} \end{cases} \]

The term \( \text{rtn}(A,B) \) represents the LOD relatedness measure which calculates the closeness of the attributes of items in different domain. Here, the bag of words i.e the determinative words of the attributes are taken from LOD which occupies megabytes of memory for storage and processing.
B. ILLUSTRATIVE RELATEDNESS CALCULATION

To inspect the performance of the proposed semantic relatedness, sample calculation is illustrated here for the pair of words (spoon, ice cream). The determinative words are discovered from LOD for the pair of words (A, B) i.e. (spoon, ice cream) as follows:

\[ X = \{\text{Serve, Eat, Utensil, Cuddle}\} \quad Y = \{\text{Frozen Food, Eat, Health, flavor}\} \]

LOD repository contains words and their relationships as a graph database. It was built based on web pages and web links. If any two words were taken, it traverses the graph and fetches all the associated information. Likewise, the determinative words were gathered which are a synonym, hyponym, hypernym and so on. Here, the utensil is the synonym of a spoon, so wherever the word spoon appears, then in all the pages of web link the word utensil will also appear. In short, hits of (utensil, spoon) are equal to 1. Likewise, the words eat and health appear in both words, so they are the co-occurrence determinative words.

The co-occurrence factor for the determinative is equal to 2. In the following Table 1, determinative words and their hits for the pair of words (spoon, ice cream) are given. In Table 1, * indicates the common determinative words, and the rest of them indicates the related value and not the actual number of pages of web link where spoon and utensil are co-ocurred. Hence spoon and utensil are synonyms, as hit number we took the maximum hits for the spoon on the determinative words. Based on the equations mentioned before, the relatedness computed between Spoon and Ice cream is 0.741. Relatedness measure value ranges from 0 to 1. Likewise, relatedness is computed for item attributes across domains which include in the MF process which helps in resolving the user cold start problem in CDRS. The CD-SemMF model suggests relevant items to new users efficiently.

C. REGULARIZATION THROUGH RELATEDNESS FORECASTING

The cross-domain semantic relatedness-based matrix factorization (CD-SemMF) model is proposed based on the hypothesis that latent vectors of interrelated items give the semantic relatedness information between items along with user preferences. The proposed model not only predicts the preferences of users \( r_{ua} \approx (p_u^T q_a) \) but also assesses the inter-domain semantic relatedness of items \( SR(u,v) = (q_u, q_v) \), where \( i \in I_s \) items belong to the source domain, \( j \in I_t \) items belong to the target domain and \( SR \) represents semantic relatedness. Therefore, the proposed model factorizes both the inter-domain item relatedness matrix and item rating matrix which links the target and source domain. Assume that \( U = U_s \cup U_t \) be the set of all users combined from source and target domain. In the proposed model the user overlaps between domains. Similarly, \( I = I_s \cup I_t \) be the set of all items combined from both domains, but they are not overlapped. Now the model learns the latent vector \( p_u \in \mathbb{R}^k \) for each user \( u \in U \) and represents each item separately from both target domain \( q_j \) with \( j \in I_t \) and source domain \( q_i \) with \( i \in I_s \) as follows:

\[
\mathcal{L}(P, Q_T, Q_s) = \sum_{i \in I_s} \sum_{j \in I_t} r_{ij}^2 - (p_i^T q_j)^2 \\
+ \lambda_c \sum_{i \in I_s} \sum_{j \in I_t} (SR_{ij} - (q_i, q_j))^2 + \lambda \left( \|P\|^2 + \|Q_T\|^2 + \|Q_s\|^2 \right) \tag{2}
\]

Where \( Q_T \) and \( Q_S \) are the matrices of target and source domains respectively with item latent vectors in rows. As seen in equation (2), summation iterates all items \( a \in I \) from source and target domain as it is necessary to simultaneously factorize the matrices of user-item preferences. The parameter \( \lambda_c \) denotes the cross-domain regularization which is used to control the contribution of inter-domain semantic relatedness. Items will have too similar latent vectors if the value of \( \lambda_c \) is large; else it transfers the knowledge limitedly between domains if \( \lambda_c \) has a low value. Here \( \lambda \) is used to avoid overfitting. Like other standard matrix factorization methods, the proposed model is trained using Alternating Least Squares. Initially, \( Q_T \) and \( Q_S \) are fixed and each \( p_u \) is solved analytically by setting the value of gradient to 0. This is because user factors are not appearing on \( \lambda_c \). The same results were obtained for the baseline Matrix Factorization model.

\[
P_u = \left( Q^T C_u Q + \lambda I \right)^{-1} Q^T C_u r_u \tag{3}
\]

To make the notation simpler, the matrix \( Q \) is defined as the row-wise concatenation of matrices \( Q_S \) and \( Q_T \), where \( C \) is the diagonal matrix with \( C_{aa} \) is the confidence value for all \( a \in I \), \( r_u \) is the vector containing user preferences for all items. Finally, the item factor \( Q_T \) and user factor \( P \) of the target.
domain are fixed and also computes the optimal solution for the item factor of the source domain. Then by setting the resultant gradient value to zero and by solving analytically, the following equation is obtained.

$$q_i = \left( \mathbf{P}^T \mathbf{C}^i \mathbf{P} + \lambda \mathbf{Q}_{I}^T \mathbf{Q}_{I} + \lambda \mathbf{I} \right)^{-1} \left( \mathbf{P}^T \mathbf{C}^i \mathbf{r}_j + \lambda \mathbf{Q}_{I}^T \mathbf{SR}_j \right) \quad (4)$$

As mentioned earlier, $r_i$ is the vector of item preferences and $SR$ denotes the $j^{th}$ row of semantic relatedness of the inter-domain matrix. As mentioned earlier $\mathbf{P}$ and $\mathbf{Q}_{I}$ are fixed to calculate the optimal value of item latent vectors of the target domain as follows:

$$q_j = \left( \mathbf{P}^T \mathbf{C}^j \mathbf{P} + \lambda \mathbf{Q}_{S}^T \mathbf{Q}_{S} + \lambda \mathbf{I} \right)^{-1} \left( \mathbf{P}^T \mathbf{C}^j \mathbf{r}_j + \lambda \mathbf{Q}_{S}^T \mathbf{SR}_j \right) \quad (5)$$

The optimal value computations are parallelized with each step, but more number of items is to be considered. So, additional steps are required for the source domain which in turn increases the training time than baseline matrix factorization approaches. To overcome such drawbacks, fast RR1 training algorithm is adapted for Alternating Least Square. Since the user factor computation equals the original matrix factorization approach, the process remains the same for P-step. From equations 2 and 4, it is noted that the Q-step of the source domain states that the extra expressions that originate from inter-domain semantic relatedness are considered just as user preferences as mentioned below:

For each item $i$ from the source domain

1. Like the baseline matrix factorization model, examples are generated for every rating $r_{ui}$
2. For every item $j \in I_T$ of the target domain
   • An input example $x_i := q_i$ is produced.
   • Relatedness is used as the dependent variable, $y_j := SR_{ij}$.
   • Constant confidence value is used, $c_{ij} := \lambda c$.
   • The factor to optimize is $w := q_i$

The above-mentioned steps help to generate the similarity terms of equation 4, which is described by the confidence matrix as $C := \lambda c I$. At the same time, the steps involved for the target domain Q-step are the same as that of the source domain.

Finally, the model works like this, if a new user came to a target domain his/her preferences are taken from the source domain. Inter-domain relatedness among items are calculated using LOD relatedness and it is added to the matrix factorization process. Then it is regularized and provides personalized items. That’s why in the mapping process LOD and MF is integrated, and recommends items to cold start user.

IV. RESULTS AND DISCUSSION

Initial processing is to compare the proposed semantic relatedness metric with other conventional metrics to prove its efficiency and later it is used in cross domain recommendations. The experimental analysis demonstrates that the proposed relatedness measure gives better results than conventional measures for the Facebook dataset in the domain of movies, music, and books. Later, ranking precision is computed for the recommendation generated by the proposed model is evaluated. It is observed that based on the source and target domain involved, the proposed model generates more accurate recommendations than baseline approaches in severe cold start scenarios.

The proposed method was also applied in the real-time amazon datasets in the domain of movies, music, and books to prove its efficiency. The proposed technique is implemented using java language and evaluated on a core i5 3.1 GHZ processor with 8GB ram and 1 TB SSD. Metrics used in evaluations, Baseline methods for efficiency comparison, datasets, and preprocessing works are discussed in detail in the following sub-sections.

A. FACEBOOK DATASET

Initially, the dataset consists of a huge set of user likes on Facebook for different items. With the help of Facebook Graph API, user likes are retrieved as 4-tuple, namely, the identifier, name of the item, liked items’ category, and timestamp of the like created. For instance, {754778727942843, The Martian, Movie, 2015-10-08T12:35:08+0000}. Here, “754778727942843” is the identifier, “The Martian” is the name of the item, Movie is the item category, and 2015-10-08T12:35:08+0000 is the timestamp when the item is liked. The item name will be created by the user who opened the Facebook page for that particular item.

Hence, for the same item, it is possible for the existence of different names. For instance, The Martian, Martian: The Movie, The Martian – Mars Film series, etc. Therefore, a user may show his/her likes for the same item on different Facebook pages. In an aim to unite and merge the items extracted from Facebook likes, the method is developed which automatically maps the name of the item with unique URIs of the equivalent DBpedia entities (LOD knowledge base). For example, unique URI: http://dbpedia.org/resource/The_Martian mapped for the item: The Martian.

1) MAPPING ITEMS WITH ITS EQUIVALENT DBPEDIA ENTITIES FOR FACEBOOK DATASET

Provided an item, entities of DBpedia which are labeled as same as item name are identified first. For that, the SPARQL query is launched on the subject of triplets with the property rdfs:label and item title as an object. For instance, the following SPARQL query is executed for The Chronicles of Narnia 2:
TABLE 2. ITEM TYPES AND ITS DBPEDIA/YAGO CLASSES CONSIDERED FOR BOOKS, MOVIES AND MUSIC DOMAINS OF FACEBOOK DATASET

| Item Type | DBpedia/YAGO classes |
|-----------|-----------------------|
| Books     |                       |
| Book      | dbo:Book,             |
|           | yago: “Book106410904”,|
|           | yago: “Book106589574” |
| Genre     | yago: “Children’s-Literature” |
| Writer    | yago-id: “alexander-mccall-smith” |
| Movies    |                       |
| Movie     | dbo:Film, yago: “Movie106613686” |
| Genre     | dbo: “MovieGenres”    |
| Director  | yago: “MovieDirector1024431418” |
| Actor     | yago: “Actor185344”   |
| Music     | dbo:Music             |
| Composition | yago: “Composer993501610” |
| Album     |                       |
| Musician  |                       |
| Genre     |                       |

To solve the ambiguity of names that equals the items of various domains, state the item type in each case that is needed to retrieve. In specific, the above-mentioned query contains the triplet clause with the property dbo:type or rdf:type. From the above-mentioned example, the subject “The Chronicles of Narnia 2” refers to the item type “film” which belongs to the class “dbo:film” in DBpedia. Item types were set based on the category of that item given in Facebook and their equivalent DBpedia entities. Their YAGO classes were discovered manually by the values of rdf:type of several entities. A list of items with their types and DBpedia/YAGO classes that are considered are given in Table 2 for three different domains, namely, music, movies, and books. Moreover, by executing the above query template, it is noted that several items are not linked to the entities of DBpedia. This is because the labels correspond to the Wikipedia redirection webpage. In such cases, the property dbp:wikiPage redirects are used by the query to attain the appropriate entities. The query executed for the movie “The Chronicles of Narnia 2” and its result is

```
SELECT DISTINCT ?item WHERE
{
  ?item rdfs:label ?name.
  FILTER regex (?name, "The.*Chronicles of Narnia. 2", "i")

  UNION
  {
    ?target rdf:type dbp:Film.
    FILTER regex (?name, "The.*Chronicles of Narnia. 2", "i")
  }
}
```

The query results in all the pairs of property-value from DBpedia of dbr:The_Chronicles_of_Narnia: Prince Caspian entity. Though, the main objective of the research is to
recommend relevant items in a cold start scenario, the model exploits only the metadata that is relevant to narrate the common preferences of dissimilar users. The returns of the query are filtered in each domain by certain properties. Those properties have high impact on the recommendation process that’s why it is considered in processing. Testing was done with all the received properties but it gives the same result as prominent properties. There is no significant improvement in the result with these additional properties instead it consumes more memory space and time. These properties were identified by factorizing the matrix. Table 3 presents the list of selected DBpedia properties for each domain of the Facebook dataset. For example, the item movie has the director, movie genres, and others as metadata. As shown in Table 3, items and their relationship represent the semantic network that is acquired automatically from the DBpedia. The statistics of the Facebook dataset considered for three preferred domains such as music, movies, and books are present in Table 4. It contains the number of items and users present in the Facebook dataset and also the number of likes given for each item. The sparsity ratio is estimated based on the total number of users, items present against the number of likes (likes/users*items), and the average number of items per user is given along with the average number of users per item in Table 4. It states that some or more users may show their interest in more than a single domain. Thus, user overlap exists between domains in this cross-domain recommender system. The number of users overlaps between domains for the Facebook dataset is shown in Table 5 with an overlap ratio.

| Table 3. Item metadata from DBpedia for books, movies and music domains |
|---------------------------------------------------------------|
| **Item Type** | **DBpedia/YAGO classes** |
|----------------|--------------------------|
| Item-genre     | dbo: “genre”             |
| Book-genre     | dbo: “LiteraryGenres”    |
| Music-genre    | dbo: “musicFusionGenre”, dbo: “stylisticOrigin” |
| Movie-genre    | dbo: “filmgenre”         |
| Item-author    | dbo: “creator”           |
| Movie-director | dbo: “director”          |
| Composer-musician | dbo: “musicComposer”, dbo: “associatedMusicalArtist” |
| Item-Item      | dbo: “basedOn”, dbo: “SubsequentWork”, dbo: “bandMember” |

| Table 4. Statistics of the Facebook dataset used for experimental purpose |
|---------------------------------------------------------------|
| **Dataset** | **Books** | **Movies** | **Music** |
|---------------|------------|------------|------------|
| Users         | 1769       | 29,843     | 45,369     |
| Items         | 3057       | 4201       | 5448       |
| Likes         | 40,869     | 9,76,501   | 26,84,462  |
| Sparsity      | 99.9924    | 99.9922    | 99.981     |
| Avg_Items/User| 23.1       | 32.72      | 59.17      |
| Avg_Users/item| 13.36      | 232.44     | 492.74     |

| Table 5. The ratio of user overlap between domains |
|---------------------------------------------------------------|
| **Source** | **Books** | **Movies** | **Music** |
| Books    | 1769       | 1395       | 1419       |
| %        | 78.85      | 80.21      |            |
| Movies   | 2195       | 29843      | 24710      |
| %        | 7.35       | 82.80      |            |
| Music    | 1214       | 24710      | 45369      |
| %        | 2.67       | 100.0      |            |
3) SEMANTICALLY ENHANCED ITEM PROFILES FOR FACEBOOK DATASET

It is fixed that the target item to be recommended is movies, books, and music. Here the obtained item metadata from LOD are distinguished into three categories as follows:

- **Features**, corresponds to the item-attribute entities are related to the considered item types in Table 3, which are distinct to the target item entities. For example, the director and genres of a specific movie.

- **Associated items**, corresponds to the item-item property of Table 3 which derives associated entities. For example, on what basis the movie is? (dbo:basedOn property), or on which band the musicians are belonging to? (dbo:bandMember property), or what the sequel of the particular movie is? (dbo:subsequentWork properties).

- **Extended features**, corresponds to the attribute-attribute property which generates extended attributes of items and does not emerging originally as metadata. For example, the fusion of particular music (dbo:musicFusionGenre property).

Semantically enriched item profiles are present in the above-mentioned item metadata types which are used in the proposed recommendation model. They are different from widely used item profiles of content-based methods. Those item profiles are made up of plain attributes. From experimental analysis, it is noted that the results attained from semantically enriched item profiles are far better than by using only plain item attributes which is given in table 8.

B. EVALUATION METHODOLOGY AND METRIC

The performance of the proposed recommendation model is validated using the improved user-based 5-fold cross-validation method [44] for user cold-start scenarios. The main aim is to study how the proposed model performs when the number of likes increases in the target domain. For that, users are roughly divided into five equal subsets. Then in each cross-validation phase, data from four subsets are kept as a training dataset and the remaining subset as the test dataset. Finally, for each user \( u \) in the test subset, their likes are roughly divided into three subsets.

1. Initially, training data was filled with user likes and iteratively down-sampled discarding one at a time to simulate the sizes of various cold start profiles.
2. To tune the parameter, a set of likes contained in validation data are used.
3. To estimate performance metrics, testing data is used.

Based on cross-domain scenarios, the above steps are modified by expanding the training subset fully with a set of likes from supporting domains to acquire actual training data for the recommendation. The recommendation model is developed for each cold-start profile size using the resultant training data. Then for every user in the testing subset, top-ranked 10 items are suggested from the target domain which is present in the training subset. Those recommended items are not known earlier by the user. Then the performance of the model is evaluated from recommendations and test data using rank-based metrics. It is observed from the results that the items suggested after the top 10 positions are irrelevant while computing the metric. Regarding the evaluation metrics, the ranking accuracy of generated item recommendations is validated using Mean Average Precision (MAP). It calculates the mean of average precision scores of items presented on top of the recommendation list. Some other evaluation metrics such as Mean Reciprocal Rank (MRR) are also examined and discovered that this metric provides the same results as MAP but the computation part takes excess time. Hence, the MAP metric is used for evaluation. MAP metric evaluates the recommended items produced by the proposed CD-SemMF model for new users. Mean Average Precision is shown in equation 6 for better understanding.

\[
\text{MAP} = \sum_{U \subseteq X} \frac{\text{AP}_U @ N}{|X|} \tag{6}
\]

Here, \(|X|\) represents the number of test sets, \(U\) represents the set of all users and \(\text{AP}_U @ N\) represents the average precision of the recommendations at Top-N of all users.

C. BASELINE METHODS FOR COMPARISON

The performance of the proposed Cross-Domain Semantic Relatedness based Matrix Factorization (CD-SemMF) model is compared with the existing baseline models, namely, CD-iMF and CD-FISM.

- **CD-iMF**: It’s a MF algorithm for positive-only feedback is trained by fast ALS.
- **CD-FISM**: The variant FISMauc is optimized for item ranking problems. Its implementation is used in the recommendation model which is accessible from LibRec [45]. FISM stands for Factored Item Similarity Model.

The optimal values for parameters of three different domains namely, Music, Movie, and Book are obtained by applying the Bayesian Optimization Technique [46] which is represented in table 6. This optimization rules out the probability of carrying out a grid search for preeminent values. Bayesian trains the model to forecast the candidate values which maximizes the given function and also reduces the uncertainty over the unknown parameter values. Here \(\lambda'\) is the dimensionality of the latent factor, \(\gamma'\) is the regularization factor, \(\alpha'\) is the confidence parameter and \(\lambda'c\) is the cross-domain regularization factor. With these optimal values mentioned for each domain in Table 6, the proposed CD-SemMF model provides better recommendations for cold start users. If any of the parameter settings go beyond the fixed limit, there is no significant change in the outcome instead it consumes time and produces the same accuracy. Hence, the system works with these tuned values to provide good results. For CD-iMF parameter settings are \(k = (12, 29, 20), \lambda = (10^7, 0.623, 1)\) and \(\alpha = (7, 6, 9)\) for books, movies and music respectively. For CD-FISM \(\lambda = 0.001, n=10^5\) and reset of the settings are same like CD-iMF.
TABLE 6. OPTIMAL HYPER-PARAMETERS FOR SEMANTIC RELATEDNESS BASED MATRIX FACTORIZATION MODEL (CD-SemMF)

| Item Type | Method      | K  | Λ  | Λc |
|-----------|-------------|----|----|----|
| Books     | CD-SemMF    | 112| 0  | 1  | $10^4$ |
| Movies    | CD-SemMF    | 10 | 1  | 16 | $10^4$ |
| Music     | CD-SemMF    | 12 | 1  | 1  | 0.002 |
| Books     | CD-SemMF    | 35 | 0  | 1  | $1.6 \times 10^6$ |
| Music     | CD-SemMF    | 10 | 1  | 1  | 0.039 |
| Movies    | CD-SemMF    | 11 | 0.571 | 4 | 0.641 |

D. INTER-DOMAIN ITEM SEMANTIC RELATEDNESS

The initial phase in processing is discovering semantic relatedness among items from both domains and later the knowledge is injected into the MF process. MAP metric is used to assess the performance of existing semantic relatedness metrics with the proposed LOD-semantic relatedness metric for the Facebook dataset is shown in Table 7. It is evident from Table 7 that the performance of the proposed LOD Semantic Relatedness metric is far better than baseline approaches such as Latent Semantic Analysis [47], Explicit Semantic Analysis [48], and Milne & Witten method [49]. For instance, in the book music domain Milne & Witten relatedness measure is 0.141, LSA is 0.067, ESA is 0.113 whereas the proposed LOD semantic relatedness measure score is 0.274. So, the superior proposed LOD-Semantic Relatedness measure is incorporated with MF in the recommendation process. This proposed model is called CD-SemMF.

TABLE 7. COMPARISON OF PROPOSED LOD-SEMANTIC RELATEDNESS METRIC WITH EXISTING METRICS ON FACEBOOK DATASETS USING MAP

| Source  | Target | LSA  | ESA  | M&W  | LOD-SR |
|---------|--------|------|------|------|--------|
| Books   | Movies | 0.058| 0.098| 0.123| 0.231  |
| Movies  | Books  | 0.067| 0.113| 0.141| 0.274  |
| Music   | Books  | 0.054| 0.061| 0.077| 0.099  |
| Music   | Movies | 0.030| 0.086| 0.136| 0.184  |
| Books   | Movies | 0.010| 0.052| 0.092| 0.192  |
| Movies  | Movies | 0.013| 0.048| 0.089| 0.156  |

E. ITEM RANKING ACCURACY FOR FACEBOOK DATASET

The ranking accuracy of items suggested by the proposed recommendation model is evaluated using the MAP metric. In an aim to study the effectiveness of cross-domain variants than single-domain variants, and also to discover whether the proposed MF model outperforms other existing models in user cold-start scenarios, the accuracy of the item rankings is analyzed. The ranking accuracy for movie recommendation is presented in Table 8 with plain item profiles. For the books-movies domain CD-iMF ranking score is 0.303, CD-FISM score is 0.298 whereas the Proposed CD-SemMF ranking score is 0.319. similarly, for the music-movie domain CD-iMF ranking score is 0.302, CD-FISM score is 0.289 and the proposed CD-SemMF score is 0.312. It is evident that movie recommendations are improved by 5.01% in the Book-Movie domain and 3.1% in the Music-Movie domain. This recommendation is done using normal item profiles which is less accurate than the semantically enriched one which is discussed next. Similarly, for other domains also the accuracy is less with normal item profiles in recommendations. Comparisons from table 9-11 is done using semantically enriched item profiles which provides better results than consuming plain item profiles in the recommendation process. Table 9 presents the ranking accuracy of recommendations generated for the book domain through Mean Average Precision. Performance of the recommendation model increases with an increase in each like. Likewise, user overlap between different domains also helps to generate more relevant recommendations. For the movies-books domain CD-FISM ranking score is 0.265, CD-iMF score is 0.284 whereas the Proposed CD-SemMF ranking score is 0.296. similarly, for the music-book domain CD-FISM ranking score is 0.258, CD-iMF score is 0.279 and the proposed CD-SemMF score is 0.292. From table 9, it is evident that book recommendations are improved by 4.45% in the Music-Book domain and 4.05% in the Movie-Book
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3175566, IEEE Access

domain. The average result reported for cold-start user profiles whose likes range from 6-10 is stable. It is observed based on the evaluation methodology, that the number of test users remains the same despite the size of the profile and also controls it by iteratively down sampling the profiles of training subsets. Similarly, Table 10 presents the ranking accuracy estimated for movie recommendations. For the books-movies domain CD-iMF ranking score is 0.393, CD-FISM score is 0.332 whereas the Proposed CD-SemMF ranking score is 0.419, similarly, for the music-movies domain CD-iMF ranking score is 0.392, CD-FISM score is 0.234 and the proposed CD-SemMF score is 0.405.

It is evident from table 10 that the movie recommendations are improved by 6.205% in the Book-Movie domain and 3.21% in the Music-Movie domain. Existing cross-domain models provide items for a new user of the movie with less accuracy than the proposed CD-SemMF. If the supporting cross-domain data contains book preferences, then the proposed MF model outperforms the best existing single-domain baseline approaches. In such a scenario, CD-SemMF generates more relevant recommendations than other popular approaches for profile size ranges from 1-10. This is owing to the high level of user overlap existing between the domains movie and book which allows CD-SemMF to calculate accurate item relatedness based on like patterns. In brief, user preferences on music and book domains are useful to generate movie recommendations for cold-start users. While the proposed recommendation model is more efficient than others by exploiting supportive likes of music. Similarly, the ranking accuracy results for music recommendation generations are presented in table 11. The music recommendations are generated for cold-start users, and the proposed model is capable enough to enhance the quality of item ranking with the help of cross-domain item metadata.

For the books-music domain CD-iMF ranking score is 0.469, CD-FISM score is 0.347 whereas the Proposed CD-SemMF ranking score is 0.506. similarly, for the movies-music domain CD-iMF ranking score is 0.486, CD-FISM score is 0.384 and the proposed CD-SemMF score is 0.513. It is evident that music recommendations are improved by 7.31% in the Book-Music domain and 5.26% in the Movie-Music domain.

### Table 8. Ranking Accuracy for Movie Recommendations in Terms of MAP on Facebook Datasets Using Plain Item Profiles

| Domain   | Method    | Number of Movie Likes |
|----------|-----------|-----------------------|
|          |           | 0  | 1  | 2  | 3  | 4  | 5  | 6-10  |
| Books    | CD-FISM   | 0.203 | 0.211 | 0.233 | 0.243 | 0.267 | 0.286 | 0.298 |
|          | CD-iMF    | 0.217 | 0.222 | 0.245 | 0.257 | 0.270 | 0.291 | 0.303 |
|          | CD-SemMF  | **0.228** | 0.234 | 0.253 | **0.271** | **0.288** | **0.301** | **0.319** |
| (proposed) |          |      |      |      |      |      |      |      |
| Music    | CD-FISM   | 0.195 | 0.209 | 0.228 | 0.240 | 0.255 | 0.267 | 0.289 |
|          | CD-iMF    | 0.201 | 0.214 | 0.231 | 0.251 | 0.264 | 0.276 | 0.302 |
|          | CD-SemMF  | **0.219** | 0.232 | 0.248 | **0.265** | **0.279** | **0.296** | **0.312** |
| (proposed) |          |      |      |      |      |      |      |      |

### Table 9. Ranking Accuracy for Book Recommendations in Terms of MAP on Facebook Datasets Using Semantically Enriched Item Profiles

| Domain   | Method    | Number of Book Likes |
|----------|-----------|-----------------------|
|          |           | 0  | 1  | 2  | 3  | 4  | 5  | 6-10  |
| Movies   | CD-FISM   | 0.235 | 0.238 | 0.245 | 0.251 | 0.256 | 0.261 | 0.265 |
|          | CD-iMF    | 0.261 | 0.264 | 0.268 | 0.272 | 0.275 | 0.279 | 0.284 |
|          | CD-SemMF  | **0.273** | 0.276 | 0.279 | **0.281** | **0.286** | **0.291** | **0.296** |
| (proposed) |          |      |      |      |      |      |      |      |
| Music    | CD-FISM   | 0.232 | 0.234 | 0.239 | 0.244 | 0.249 | 0.253 | 0.258 |
|          | CD-iMF    | 0.252 | 0.258 | 0.262 | 0.264 | 0.266 | 0.271 | 0.279 |
|          | CD-SemMF  | **0.263** | 0.267 | 0.272 | **0.276** | **0.281** | **0.285** | **0.292** |
| (proposed) |          |      |      |      |      |      |      |      |
### Table 10. Ranking Accuracy for Movie Recommendations in Terms of MAP on Facebook Datasets Using Semantically Enriched Item Profiles

| Domain | Method        | Number of Movies Likes |
|--------|---------------|------------------------|
|        |               | 0         | 1         | 2         | 3         | 4         | 5         | 6-10  |
| Books  | CD-FISM       | 0.248     | 0.257     | 0.273     | 0.283     | 0.297     | 0.307     | 0.332 |
|        | CD-iMF        | 0.267     | 0.288     | 0.315     | 0.337     | 0.355     | 0.367     | 0.393 |
|        | CD-SemMF      | **0.328** | **0.344** | **0.358** | **0.371** | **0.382** | **0.402** | **0.419** |
|        | (proposed)    |           |           |           |           |           |           |       |
| Music  | CD-FISM       | 0.105     | 0.119     | 0.133     | 0.159     | 0.191     | 0.217     | 0.234 |
|        | CD-iMF        | 0.301     | 0.314     | 0.331     | 0.351     | 0.363     | 0.375     | 0.392 |
|        | CD-SemMF      | **0.317** | **0.321** | **0.346** | **0.362** | **0.377** | **0.391** | **0.405** |
|        | (proposed)    |           |           |           |           |           |           |       |

### Table 11. Ranking Accuracy for Music Recommendations in Terms of MAP on Facebook Datasets Using Semantically Enriched Item Profiles

| Domain | Method        | Number of Music Likes |
|--------|---------------|-----------------------|
|        |               | 0         | 1         | 2         | 3         | 4         | 5         | 6-10  |
| Books  | CD-FISM       | 0.153     | 0.184     | 0.205     | 0.236     | 0.276     | 0.318     | 0.347 |
|        | CD-iMF        | 0.201     | 0.268     | 0.321     | 0.353     | 0.391     | 0.424     | 0.469 |
|        | CD-SemMF      | **0.312** | **0.358** | **0.391** | **0.414** | **0.446** | **0.471** | **0.506** |
|        | (proposed)    |           |           |           |           |           |           |       |
| Movies | CD-FISM       | 0.193     | 0.216     | 0.259     | 0.287     | 0.322     | 0.351     | 0.384 |
|        | CD-iMF        | 0.293     | 0.336     | 0.378     | 0.403     | 0.424     | 0.454     | 0.486 |
|        | CD-SemMF      | **0.361** | **0.393** | **0.426** | **0.438** | **0.465** | **0.487** | **0.513** |
|        | (proposed)    |           |           |           |           |           |           |       |

### F. Performance Evaluation Using Amazon Dataset

Amazon dataset is chosen for real-time evaluation. The following figure 2 presents the format of the Amazon dataset [50]. It contains a lot of item information like title, category of the item, id, ASIN No. and so on. Among them all, only a very few data are required for the experimental purpose. They are id, rating, customer, and group which represents the customer id, item id, numerical rating value, and group type provided by the customer of that particular item. But, there are items with no user rating. For example, in figure 2 the id: 9 has no rating. Such types of entry are irrelevant and so they are removed from the dataset. It is also filtered out for accurate recommendations and converts the data into four tuples with id, group, customer, and rating. The overall description of the Amazon dataset taken for the experiment is given in Table 12. It contains a category of the item, ratings for each item, the number of items, and users present in the dataset.

### Table 12. The Statistics of Amazon Dataset

| Category | Ratings | Items | Users |
|----------|---------|-------|-------|
| Book     | 4,591,303 | 393,561 | 984,852 |
| Music    | 1,388,490 | 103,144 | 460,524 |
| Movies   | 813,487   | 19,828  | 239,056 |
A similar evaluation process or setting is followed for the amazon dataset such as mapping of amazon data with DBPedia entities, semantically annotated dataset, enhanced item profiles, and 5-fold cross-validation. Evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are used to validate the recommendation accuracy of the proposed CD-SemMF model with existing recommendation models, namely, CD-FISM and CD-iMF. Tables 13 to 15 presents the MAE and RMSE value for the shared domain of proposed and baseline models for comparative analysis. Calculation of MAE and RMSE is shown in equations 7 and 8.

\[
MAE = \frac{1}{n} \sum_{(u,i) \in T} | \hat{r}_{ui} - r_{ui} | \tag{7}
\]

\[
RMSE = \sqrt{\frac{1}{n} \sum_{(u,i) \in T} (\hat{r}_{ui} - r_{ui})^2} \tag{8}
\]

Here “\(r_{ui}\)" represents actual value, “\(\hat{r}_{ui}\)” is predicted value and ‘\(n\)’ is the number of observations. Table 13 indicates that MAE and RMSE values are lesser in the proposed model than the existing models. RMSE of CD-iMF model for the music-book domain is 1.192 whereas for the proposed model is 1.064. Similarly, RMSE of CD-iMF model for the movie-books domain is 1.173 whereas for the proposed CD-SemMF is 1.045. Thus, by including LOD-semantic relatedness and item metadata in Matrix Factorization processing, the performance of the CD-SemMF model increases by 12.03% for the Music-Book domain and 12.24% for the Movies-Book domain. For movie recommendations shown in Table 14, the MAE and RMSE values for the proposed CD-SemMF model are 0.738, 1.021 for the Book-Movie domain, and 0.764, 1.054 for the Music-Movie domain, respectively. Whereas for the baseline models like CD-FISM and CD-iMF the error rate is higher than the proposed model. RMSE of CD-iMF model for the book-movies domain is 1.184 whereas for the proposed model is 1.021. Similarly, RMSE of CD-iMF model for the music-movies domain is 1.234 whereas for the proposed CD-SemMF model is 1.054. Recommendation performance of the proposed model increases by 15.96% for the Book-Movie domain and 17.07% for the Music-Movie domain. For music recommendations shown in Table 15, the MAE and RMSE value for the proposed CD-SemMF model is 0.750, 1.039 for the Book-Music domain, and 0.710, 0.997 for the Movie-Music domain, respectively. Whereas for the baseline models like CD-FISM and CD-iMF the error rate is higher than the proposed model. RMSE of CD-iMF model for the book-music domain is 1.156 whereas for the proposed model is 1.039. Similarly, RMSE of CD-iMF model for the movies-music domain is 1.051 whereas for the proposed CD-SemMF is 0.997. Recommendation performance of the proposed model increases by 11.26% for the Book-Music domain and 5.41% for the Movie-Music domain. Hence, the proposed CD-SemMF model resolves the user cold start issues in the collaborative filtering recommender system with the inclusion of item metadata (DBpedia) and LOD semantic relatedness in matrix factorization.


| Method       | Music-Book | Movies-Book |
|--------------|------------|-------------|
|              | MAE | RMSE | MAE | RMSE |
| CD-FISM      | 0.861 | 1.336 | 0.883 | 1.382 |
| CD-iMF       | 0.830 | 1.192 | 0.828 | 1.173 |
| CD-SemMF (Proposed) | **0.768** | **1.064** | **0.751** | **1.045** |

| Method       | Book-Movies | Music-Movies |
|--------------|-------------|--------------|
|              | MAE | RMSE | MAE | RMSE |
| CD-FISM      | 0.866 | 1.348 | 0.870 | 1.358 |
| CD-iMF       | 0.829 | 1.184 | 0.838 | 1.234 |
| CD-SemMF (Proposed) | **0.738** | **1.021** | **0.764** | **1.054** |

| Method       | Book-Music | Movies-Music |
|--------------|------------|--------------|
|              | MAE | RMSE | MAE | RMSE |
| CD-FISM      | 0.854 | 1.295 | 0.801 | 1.112 |
| CD-iMF       | 0.810 | 1.156 | 0.758 | 1.051 |
| CD-SemMF (Proposed) | **0.750** | **1.039** | **0.710** | **0.997** |

V. Conclusion and Future Work

The most considered and well accepted solution for user cold start issues in collaborative filtering algorithms is cross-domain. CF only mines the pattern of user-item preferences and ignores item content information to connect interested domains. While some other algorithms exploit content-based relations to effectively connect domains of interest. A recent initiative called Linked Open Data (LOD) provides large interconnected repositories of structured knowledge which is used in a way to relate different types of data. This type of heterogeneity helps to launch content-based links between different types of items. Therefore, in cross-domain recommendations, LOD can be used to connect different domains. Hence, in the proposed work Linked Open Data were exploited to extract the item metadata which helps for better recommendation across domains. With this additional information, it is feasible to discover the relation between items of various domains and eventually easy to compute inter-domain item relatedness. A novel Cross-Domain Semantic relatedness-based Matrix Factorization model (CD-SemMF) is proposed for recommendations that exploit the computed relatedness between items to link knowledge across different domains. From the experimental analysis in the severe user cold start scenarios, it is evident that the cross-domain recommendation algorithm generates accurate suggestions for new users in the target domain by exploiting item metadata based on the source and target domains involved in the recommendation process. However, the priority of the CF recommender system in cold start scenarios is to engage the target user by generating relevant suggestions than diverse or non-relevant items. The model proposed in this research work belongs to the knowledge transfer cross-domain recommender system. The proposed approach is applied to the linked-domain exploitation task which addresses the user cold start issue. Moreover, it also inferred from the experimental results that item metadata retrieved from DBpedia and the relatedness measure are the added advantage of recommendations with minimum user overlap. The proposed CD-SemMF model provides better recommendations for new users than existing models. In the future, the work represented in this article will be extended by inheriting deep learning techniques in cross domain recommendations and also improvement should be done in the experimental section by considering more cross domain datasets linked to DBpedia.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENT

The authors gratefully acknowledge the Science and Engineering Research Board (SERB), Department of Science & Technology, India, for financial support through Mathematical Research Impact Centric Support (MATRICS)
scheme (MTR/2019/000542). The authors also acknowledge SASTRA Deemed University, Thanjavur, for extending infrastructural support to carry out this research work.

REFERENCES
[1] I. Fernández-Tobías, I. Cantador, P. Tomeo, V. W. Anelli and T. D. Noia, “Addressing the user cold start with cross-domain collaborative filtering: exploiting item metadata in matrix factorization,” User Modeling and User-Adapted Interaction, vol. 29, pp. 443-486, Jan. 2019.
[2] D. Veras, R. Prudêncio, and C. Ferraz, “CD-CARS: Cross-domain context-aware recommender systems,” Expert Systems with Applications, vol. 135, pp. 388-409, Nov. 2019.
[3] N. Shadbolt, T. Berners-Lee, and W. Hall, “The semantic web revisited,” IEEE intelligent systems, vol. 21, pp. 96-101, Feb. 2006.
[4] C. Bizer, T. Heath, and T. Berners-Lee, “Linked data: The story so far,” In Semantic services, interoperability and web applications: emerging concepts, 2011, pp. 205-227.
[5] Y. Lin, H. Wang, J. Chen, T. Wang, Y. Liu, H. Ji, Y. Liu and P. Natarajan, “Personalized Entity Resolution with Dynamic Heterogeneous Knowledge Graph Representations,” arXiv preprint arXiv:2104.02667, Apr. 2021.
[6] T. D. Noia, V. C. Otsuni, P. Tomeo, and E. D. Sciascio, “Sprank: Semantic path-based ranking for item recommendations using linked open data,” ACM Transactions on Intelligent Systems and Technology, Vol. 8, pp. 1-34, Jan. 2017.

F. Shapira, L. Rokach, and S. Freilikman, “Facebook single and cross domain data for recommendation systems,” User Modeling and User-Adapted Interaction, Vol. 23, pp. 211-247, 2013.
[11] H. Hwangbo and Y. Kim, “An empirical study on the effect of data sparsity and data overlap on cross domain collaborative filtering performance,” Expert Systems with Applications, Vol. 89, pp. 254-265, 2017.
[12] S. Gao, H. Luo, D. Chen, S. Li, P. Gallinari, and J. G. Guo, “Cross-domain recommendation via cluster-level latent factor model,” In Joint European conference on machine learning and knowledge discovery in databases, Sep. 2013, pp. 161-176.
[13] S. Sahebi, and P. Brusilovsky, “Cross-domain collaborative recommendation in a cold-start context: The impact of user profile size on the quality of recommendation,” In International Conference on User Modeling, Adaptation, and Personalization, Jun. 2013, pp. 289-295.
[14] S. Wei, X. Zheng, D. Chen, and C. Chen, “A hybrid approach for movie recommendation via tags and ratings,” Electronic Commerce Research and Applications, Vol. 18, pp. 83-94, 2016.
[15] M. M. Khan, R. Ibrahim, and H. Ghani, “Cross-domain recommender systems: a systematic literature review,” ACM Computing Surveys (CSUR), Vol. 50, pp. 1-34, 2017.
[16] R. P. Karumur, T. T. Nguyen, and J. A. Konstan, “Exploiting user interests for user modeling on Twitter for link recommendations,” In proceedings of the 12th international conference on semantic systems, Sep. 2016, pp. 81-88.
[17] O. Goga, H. Lei, S. H. K. Baranshahrati, G. Friedland, R. Sommer, and R. Teixeira, “Exploiting innocuous activity for correlating users across sites,” In Proceedings of the 22nd international conference on World Wide Web, May. 2013, pp. 447-458.
[18] J. Ma, Y. Qiao, G. Hu, Y. Huang, M. Wang, A. K. Sangaiah, and Y. Wang, “Balancing user profile and social network structure for anchor link inferring across multiple online social networks,” IEEE Access, Vol. 5, pp. 12031-12040, 2017.
[19] J. Protasiewicz, W. Pedrycz, M. Kozłowski, S. Dadas, T. Stanisławek, A. Kopacz, and M. Gałęziowska, “A recommender system of reviewers and experts in reviewing problems,” Knowledge-Based Systems, Vol. 106, pp. 164-178, 2016.
[20] J. Liu, J. Shi, W. Cai, B. Liu, W. Pan, Q. Yang, and Z. Ming, “Transfer Learning from APP Domain to News Domain for Dual Cold-Start Recommendation,” In RecSysKTL, pp. 38-41, Aug 2017.
[21] N. Mirbakhsh, and C. X. Ling, “Improving top-n recommendation for cold-start users via cross-domain information,” ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 9, pp. 1-19, 2015.
[22] C. Wongchokkansiti, J. Peltonen, T. Ruotsalo, P. Bandypadhyay, G. Facucci, and P. Brusilovsky, “User model in a box: Cross-system user model transfer for resolving cold start problems”, In International Conference on User Modeling, Adaptation, and Personalization, Jun. 2015, pp. 289-301.
[23] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Application of dimensionality reduction to recommender systems-a case study,” Minnesota Univ Minnie Dept of Computer Science. 2000.
[24] S. Funk, “Netlix update: Try this at home”, 2006.
[25] Y. Hu, Y. Koren, and C. Volinsky, “Collaborative filtering for implicit feedback datasets”, In 2008 Eighth IEEE International Conference on Data Mining, Dec. 2008, pp. 263-272.
[26] G. Takács, and D. Tikk, “Alternating least squares for personalized ranking,” In Proceedings of the sixth ACM conference on Recommender systems, Sep. 2012, pp. 83-90.
[27] G. Alexandridis, G. Siolas, and A. Stafyllopatis, “Enhancing social collaborative filtering through the application of non-negative matrix factorization and exponential random graph models,” Data Mining and Knowledge Discovery, Vol. 31, pp. 1031-1059, 2017.
[41] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, “BPR: Bayesian personalized ranking from implicit feedback,” arXiv preprint arXiv:1205.2618. 2012.

[42] S. Kabbur, X. Ning, and G. Karypis, “Fism: factored item similarity models for top-n recommender systems,” In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, Aug. 2013, pp. 659-667.

[43] M. A. Salahli, “An approach for measuring semantic relatedness between words via related terms,” Mathematical and Computational Applications, Vol. 14, pp. 55-63, 2009.

[44] D. Kluver, and J. A. Konstan, “Evaluating recommender behavior for new users,” In Proceedings of the 8th ACM Conference on Recommender Systems, Oct. 2014, pp. 121-128.

[45] M. Mansoury, R. Burke, A. Ordonez-Gauger, and X. Sepulveda, “Automating recommender systems experimentation with librec-auto,” In Proceedings of the 12th ACM Conference on Recommender Systems, Sep. 2018, pp. 500-501.

[46] J. Snoek, H. Larochelle, and R. P. Adams, “Practical bayesian optimization of machine learning algorithms,” arXiv preprint arXiv. pp: 1206.2944, 2012.

[47] T. Hofmann, “Probabilistic latent semantic analysis”, arXiv preprint arXiv:1301.6705. 2013.

[48] S. A. Salloum, R. Khan, and K. Shaalan, “A Survey of Semantic Analysis Approaches,” In AICV, Mar. 2020, pp. 61-70.

[49] I. H. Witten, and D. N. Milne, “An effective, low-cost measure of semantic relatedness obtained from Wikipedia links”, 2008.

[50] A. K. Sahu, and P. Dwivedi, “Knowledge transfer by domain-independent user latent factor for cross-domain recommender systems,” Future Generation Computer Systems, Vol. 108, pp. 320-333, 2020.

[51] https://dbpedia.org/ontology/Film