Dimensionality Reduction Using Latent Variable across the Domains in Recommender System

Dr. C. Valliyammai, M. Nanthini, & S. Ephina Thendral
Department of Computer Technology
Madras Institute of Technology, Anna University, Chennai
cva@annauniv.edu , mnanthini.cse@gmail.com , ephina.chris@gmail.com

***Abstract*—Dimensionality reduction plays an important role in big data analytics and machine learning for the past decades. While exploring the large volumes of data, it is necessary to perform the larger computation. In order to overcome this, a novel latent variable based dimensionality reduction across the domains in Recommender System (RS) is proposed. Firstly, we define the latent class corresponding to the attributes from two domains and user profiles. Then many-to-one mapping of attributes to a latent class variable is achieved. Finally, the entire data variables are reduced to five latent class variables and sharing the knowledge across the domains. The overall dimensionality reduction is very useful for easy processing of data and reducing the processing time in various applications. Compared with the traditional dimensionality reduction method, the proposed method discovers the hidden variable from the observed variable without any loss of information.**

***Keywords*—Dimensionality Reduction, Latent Variable, Attribute Mapping, Cross-Domain Analysis, Recommender System**

I. INTRODUCTION

Big data handling constitutes one of the important challenges in RS. One way to handle big data is Dimensionality Reduction (DR). DR is the process of reducing the number of variables which are observed and preserving the original content as much as possible. DR can be divided into Feature Selection (FS) and Feature Extraction (FE) based on the application. FS is the selection of a subset of the whole dataset that represents the whole data. FE is the process of transforming the high dimensional data into lower dimensional data. With the help of these techniques, DR can effectively decrease the problem of over-fitting and "curse of dimensionality" in machine learning, data processing in big data analytics and various applications.

Nowadays, a large volume of information on the websites is in the form of user ratings, reviews, user opinions, remarks, user feedback, and comments about any item and it is difficult for the engine to search the user interest and predict their preference level. The Recommender System (RS) came into existence, supports both customers and providers in their decision-making process. RS is information filtering system that attempts to find the rating or preference that a user would give to an item. The most common recommender filtering techniques are Collaborative Filtering (CF) and content-based filtering. Collaborative recommendation [1, 2] is also known as a social recommendation, it provides recommendations based on similar people preferences, likes, dislikes, and reviews. The content-based recommendation is a cognitive filtering, it provides recommendation according to the content of an item and the user profile. Hybrid recommender systems combine the two popular approaches collaborative and content-based systems by collecting the user profiles learning the model based on the information and maintaining some useful records using an efficient information retrieval techniques and other content-based methods, and directly comparing the user profiles to find the similar users based on collaborative filtering [3, 4] and provide recommendations.

In traditional systems, filtering techniques are providing a recommendation based on single domain information. The domain is one of the characteristics of recommender system and it is a particular thought of field, activity or user interest.

Considering two domains [5, 6, 7] in RS, the numbers of attributes to be considered are larger and different compared to the single domain. DR helps to reduce the overall dimensionality in two domains and improve the prediction accuracy of user preference in RS by discovering the hidden knowledge of the whole data set.

The data available on the online websites are of various types such as continuous/discrete, numerical/categorical, structured /unstructured, etc. Continuous data is the range can take any value between its maximum and minimum value. Structured data refers to the information with the fully organized manner and can enable quick search algorithm. Unstructured data have a lack of structure which makes energy consuming task. Categorical data can take only certain values and is also known as discrete data. It can be classified as nominal or ordinal variables. One of the challenges is analyzing such variety of data and dimensionality reduction techniques can be applied for big data analysis.

The rest of the paper is organized as follows: Section 2 provides details of various dimensionality reduction techniques; Section 3 presents the latent variable based dimensionality reduction across the domains, and Section 4 presents an experimental evaluation of the proposed work with real-world trip advisor datasets. Finally, Section 5 concludes the paper.

II. RELATED WORK

Social recommender system [8] uses the CF to make personalized recommendations. Most of the CF techniques mainly use the one-dimensional clustering methods to group users and items separately. However, the one-dimensional

IRJECE

method usually neglects the necessary information in another dimension. Daqiang Zhang, et. al. uses the bi-clustering technique, that group both the user dimension and item dimension concurrently in the user-item matrix. Correspondingly, the bi-clustering method outperforms the one-way cluster method of overcoming the sparsity problem and reducing the high-dimensional matrices in recommender systems.

Music recommender system uses the CF recommendation algorithm [9] based on Single Value Decomposition (SVD). The main aim of SVD is dimensionality reduction i.e. considering high dimensional features and disintegrating it to lower dimensional features. SVD is a technique where a matrix is factored into a series of linear likeness. This likeness will uncover the hidden structure of the matrix. The SVD can be represented from three distinct viewpoints. First, SVD reconstructs a matrix of corresponding variables into interrelated one that leads to better understanding of the relationship between all data points which might not be noticeable in their original construction. Secondly, SVD is used to find the relationship between several data points in the matrix and co-ordinate the data points in the product matrix [10] so that the data points display the maximal deviation. Thirdly, the highest approximations in the original data points are determined using fewer dimensions.

Supervised dimensionality reduction is the distinct method uses the class labels corresponding to the data in order to accomplish a mapping such that the final mapped data correspond to different classes will be right separated. Dimitrios Bouzas et. al. proposed a supervised linear dimensionality reduction [11] technique. A novel method for DR uses the mutual information as the criterion between the reconstructed data and the appropriate class labels.

Mehmet Gönen proposed a Bayesian Supervised Dimensionality Reduction (BSDR) [12] for classification and information retrieval. A novel BSDR algorithm learned the projection matrix and the learning parameters together in a linear model in order to increase the prediction accuracy in the projected subspace. Gibbs sampling is used to learn the probabilistic model for multi-label classification. The deterministic variation approximation is used to reduce the computational complexity. Intrinsic dimensionality is calculated using Automatic Relevance Determination (ARD).

Randomized dimensionality reduction [13] combines both random projections and SVD for k-means clustering. Christos Boutsidis et. al. proposed a FS technique for k-means clustering and two FE methods. The first FE technique is a random projection and it outperforms the existing FE method in terms of time complexity and the overall features to be extracted. The second FE technique is a fast approximate SVD factorization and it also improves the results in terms of time complexity. The algorithms are randomized and constantly ensure the best k-means objective value.

The major problems of recommender system are data sparsity and cold start issues. In a social network, the star-structured hybrid graph is used for making recommendation across the domains and the useful information is shared from a source domain to the target domain. Various factors of item transferability, including popularity and behavioral consistency, are determined. Meng Jiang, et. al. proposed a method called Hybrid Random Walk (HRW) [14], which considers some in a target domain. Map across the useful knowledge within the domain and approximately predict the relationships between the user and item in the target domain.

Bin Li, et. al. proposed a CF techniques in cross domain [15] in order to share similar rating information across several related domains to increase the filtering performance using ROST (Ratings Over Site Time) for reducing data sparsity problem in the target domain, where several similar sites are considered as same filtering domains and the products in a target domain depend on their similarity to the same ones and ROST (time) for creating a model for dynamic changes of user interest.

The recommendation in social network gain more popularity and attaining a successful service for providing quality of service and user satisfaction. These applications enable the users to provide various implicit feedbacks and ratings in social networks through their daily interactions. The users in many social networks can like some items and provide feedbacks. With the exponential growth of online information, the sparse data gradually decrease the quality and correlation factor of feedbacks and recommendation. Fei Hao, et. al. proposed a novel Spatial Social Union (SSU) [16], it is a similarity measurement based approach and calculate the similarity between two users that combines the interconnection between users and items in addition to location information.

III. PROPOSED SYSTEM

Latent variables are derived from the domains in order to achieve DR. This section describes the latent variables, cross-domain analysis, attributes mapping from user profiles to the latent variable and domain level attributes mapping to the latent variables.

A. Latent Variables

Latent variables are the variables that are not present in the original dataset but are inferred from the variables that are present in the original dataset (observed variables). By introducing the set of latent variables, unobserved inferences from the real world datasets are discovered and uncovering the hidden patterns that lead to knowledge discovery. Mathematical models that aim to derive the latent variables from observed variables are called latent variable models.

One advantage of using latent variable is a dimensionality reduction in big data. A large number of observed variables can be combined in a model to reproduce an underlying concept, making it easier for better understanding of data. The five unobserved latent variables are categorical and are labeled as L1, L2, L3, L4 and L5.

B. Cross-Domain Analysis

A domain is a specific area of thought, action, or user interest. Many of the recommender systems concentrate on single domains and they provide recommendations within the same domain where users gave their feedback and ranks. The

IRJECE

combination of various domains into single domain make the recommender engine provides recommendation list across the domain. An algorithm used in domain analysis is able to provide a recommendation of the product in the target domain to users who gave ratings only in the source domain. For example if restaurant domain is referred as the source domain and the tourist spot domain is referred as the target domain and the knowledge is shared across restaurant domain and tourist spot domain.

C. *Various Domain Level*

- Attribute level (Comedy↔ Thriller)
  - Same kind of items, distinguish values of certain attributes
- Type-level (Movies↔ Books)
  - Distinct types, sharing some attributes.
- Item level (Movies↔ Restaurants)
  - Distinct types, differing in most, if not all attributes.
- System level (Netflix↔ Movie lens)
  - Almost the same items, collected in different ways and/or from different operators.

The domains considered for dimensionality reduction is based on the type level, sharing some attributes but are of different types. The two domain level attributes are collected along with user profiles. Figure 1 represents the dimensionality reduction across the domains by extracting common features and attribute mapping.

Common features from two domains are identified and the five latent variables are initialized. Three matrices are constructed to perform many to one mapping of attributes to the latent variables. Dimensionally reduced matrices are used for the various analyzing process in recommender system and discovering hidden knowledge.

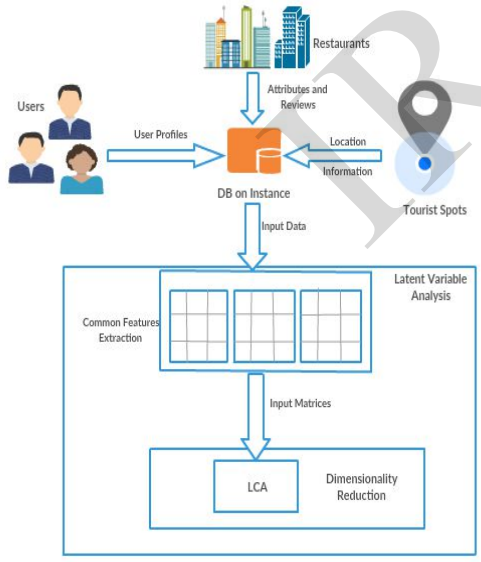


Fig. 1: Dimensionality Reduction across the Domains

D. *Attributes Mapping to Latent variables*

The proposed system uses two domain level information and user profiles for attribute mapping. The two domains considered for dimensionality reduction in RS are Restaurants and Tourist Spot. Common features across the domain can be extracted using latent class analysis. Many-to-one mapping of attributes to latent variables is done to reduce the dimensionality of overall attributes in a dataset considering related domains. The user profiles are name, location, gender, history, badges collection, places visited, travel styles, contribution, price range, date of visit, ratings and reviews.

Restaurant attributes are name, location, restaurant ratings, reviews by the user, cuisine and review count. Tourist spot attributes are name, ratings, tourist spot reviews, spot location and the type. The general equation for latent variable derivation from observed attributes is given as equation (1),

$$L = U(\sum(f(u), f(r), f(t))) \quad (1)$$

Where f(u), f(r) and f(t) are the functions for user profiles, restaurant and tourist spot respectively which are mapping from various observed variables corresponding to the latent variables.

Table I consists of various symbols used to derive latent variables and the appropriate explanation for the symbols. The symbols represent the attributes of user, restaurants and tourist spots.

TABLE I: Symbols And Explanation

Symbols	Explanation
lo_{ul}	User location
lo_{rl}	Restaurant location
lo_{tl}	Tourist Spot location
pl_{ut}	Type of places visited by the user
ra_{ur}	Rating given by the user for restaurant
ra_{ut}	Rating given by the user for tourist spot
re_{ur}	Review given by the user for restaurant
re_{ut}	Review given by the user for tourist spot
pr_{ur}	Price ranges for restaurant visited by the user
pr_{ut}	Price ranges for tourist spot visited by the user
ra_{rr}	Rating for the restaurant
ra_{tr}	Rating for the tourist spot
re_{rr}	Review for the restaurant
re_{tr}	Review for the tourist spot
pr_{rr}	Price ranges for restaurant
pr_{tr}	Price ranges for tourist spot
co_{uc}	Contribution of user to the system
ba_{ub}	Badges collected by the user
pl_{uc}	Count of places visited by the user
rc_{rc}	Review count for restaurant
rc_{tc}	Review count for tourist spot
ts_{us}	Travel styles of user
ge_{ug}	Gender of user
cu_{rc}	Restaurant cuisine
ty_{ts}	Type of tourist spot
hi_{uh}	User history
hy_{th}	Holidays of the tourist spot
dv_{ur}	Date of visit to the restaurant by the user
dv_{ut}	Date of visit to the tourist spot by the user

IRJECE

The first latent variable mapping is based on the latitude and longitude value of the location is labeled as city and outer region. The mapping function for first latent variable L1 based on location is given as equation (2),

$$L1 = \bigcup(\sum_{U}(lo_{ul}, pl_{ut}), \sum_{R} lo_{rl}, \sum_{T} lo_{tl}) \quad (2)$$

Second latent variable mapping is based on the level of the user, restaurant and tourist spot. The second variable is labeled as high, medium and low. The mapping function for second latent variable is given as equation (6),

$$f(u_{L2}) = \sum((ra_{ur}, pr_{ur}, re_{ur}), (ra_{ut}, pr_{ut}, re_{ut})) \quad (3)$$

$$f(r_{L2}) = \sum(ra_{rr}, pr_{rp}, re_{re}) \quad (4)$$

$$f(t_{L2}) = \sum(ra_{tr}, pr_{tp}, re_{te}) \quad (5)$$

$$L2 = \bigcup(f(u_{L2}), f(r_{L2}), f(t_{L2})) \quad (6)$$

Equation (3), equation (4) and equation (5) represents the mapping function for user, restaurant and tourist spot respectively based on ratings , price range and reviews. L2 is the second latent variable which is derived from $f(u_{L2})$, $f(r_{L2})$ and $f(t_{L2})$, mappings of various attributes corresponding to the latent variable L2.

Third latent variable L3 mapping is based on the popularity of the two domains and the users are labeled as high and low. The mapping function is given as equation (7),

$$L3 = \bigcup(\sum_{U}(co_{uc}, ba_{ub}, pl_{uc}), \sum_{R}(ra_{rr}, rc_{rc}), \sum_{T}(ra_{tr}, rc_{tc})) \quad (7)$$

Fourth latent variable L4 mapping is based on the type of the restaurant and the tourist spot, user might prefer based on their reviews and ratings and it is labeled as numeric values. L4 is derived from equation (8), equation (9) and equation (10), where $f(u_{L4})$, $f(r_{L4})$ and $f(t_{L4})$ represents the mapping functions for user, restaurant and tourist spot respectively. The mapping notation is given as equation (11),

$$f(u_{L4}) = \sum((ra_{ur}, re_{ur}), (ra_{ut}, re_{ut}), ge_{ug}, ts_{us}, pl_{ut}) \quad (8)$$

$$f(r_{L4}) = \sum(cu_{rc}) \quad (9)$$

$$f(t_{L4}) = \sum(ty_{ts}, lo_{tl}) \quad (10)$$

$$L4 = \bigcup((f(u_{L4}), f(r_{L4}), f(t_{L4})) \quad (11)$$

Fifth latent variable L5 mapping is based on the season and it is labeled as summer, winter and throughout the year. The attributes considered for this mapping are type of the places, restaurant type and date of visit for the location by the user. The mapping function is given as equation (12),

$$L5 = \bigcup(\sum_{U}(hi_{uh}, dv_{ur}, dv_{ut}) \sum_{R} cu_{rc}, \sum_{T}(ty_{ts}, hy_{th})) \quad (12)$$

IV. EXPERIMENTS AND RESULTS

A. Data Collection

A software tool that roams around web pages and domains by following hyperlinks to discover and access web page content is termed as a crawler. Outwit hub breaks down web pages into their different constituents. Navigating from page to page automatically, it extracts information elements and organizes them into usable collections. The contents extracted from a web page are presented in an easy and visual way. The extracted data can be exported to CSV, HTML, Excel or SQL databases, while images and documents, are directly saved to the hard disk. Trip advisor dataset contains the user profiles and the attributes to be considered from two domains.

B. Experimental setup

The experimental setup consists of 1000 records from a trip advisor [17] dataset in the experimental setup. It consists of 100 users and each user having unique profiles. The restaurant and the tourist spot attributes are collected based on the location, 100 users are visited. The user, restaurant, tourist spot variables are mapped to five latent variables.

C. Results

The experiments are conducted to obtain the latent variables L1, L2, L3, L4 and L5 with sub-labels. The results show that raw attributes are reduced to five latent variables for user profiles, restaurants and tourist spots are listed in Table II and remove the variations in the number of attributes considered for analysis. Final matrix consists of latent variables corresponding to the user, restaurants and tourist spots.

TABLE II: Dimensionality Reduction to Five Latent Variables

Domains	L1	L2	L3	L4	L5
User 1	City	Low	Low	0	Winter
User 2	Outer	High	High	1	Summer
. .					
User n	Outer	Mediu m	Low	1	Throughout
Restaurant 1	City	High	Low	1	Summer
Restaurant 2	Outer	High	High	0	Throughout
. .	. .	. .	. .	. .	. .
Restaurant n	Outer	Low	High	1	Winter
Touristspot1	City	High	Low	1	Summer
Touristspot2	City	Mediu m	High	2	Winter
. .	. .	. .	. .	. .	. .
Tourist spot n	Outer	Low	High	0	Throughout

IRJECE

V. CONCLUSIONS

The latent variable based dimensionality reduction is considered for related domains in the recommendation system. Common features across the domain are extracted using latent class analysis in order to reduce the dimensionality of big datasets. This technique is taking user profiles into consideration and two domain attributes based on the location, users visited for attribute mapping. The problem of high dimensionality in big data is addressed using latent variable mapping.

This research can be extended to reduce the data sparsity problem in the feedback matrix of recommender system using prediction techniques. The proposed recommender system can be tested in a real time distributed environment when the data size is huge.

REFERENCES

[1] Lina Yao, Quan Z. Sheng, Anne. H.H. Ngu, Jian Yu, and Aviv Segev, "Unified Collaborative and Content-Based Web Service Recommendation", in IEEE Transactions on Services Computing, vol. 8, no. 3, pp. 453-466, May/June. 2015.

[2] Mohamed Amine Chatti, Simona Dakova, Hendrik Thus, and Ulrik Schroeder, "Tag-Based Collaborative Filtering Recommendation in Personal Learning Environments", in IEEE Transactions on Learning Technologies, vol. 6, no. 4, pp. 337-349, 2014.

[3] Go, Yang J, Park H. and S. Han, "Using online media sharing behaviors implicit feedback for collaborative filtering", in IEEE International Conference on Social Computer, pp. 439–445, 2015.

[4] Enrich, M., Braunhofer, and Ricci, "Cold-Start Management with Cross Domain Collaborative Filtering and Tags", in the fourteenth International Conference on E-Commerce and Web Technologies, pp. 101-112, 2013.

[5] P. Cremonesi, A. Tripodi, and R. Turrin, "Cross-domain recommender Systems", in Proc. IEEE 11th International Conference on Data Mining Workshops (ICDMW), Vancouver, BC, Canada, pp. 496–503, 2014.

[6] Ivan Cantador, and Paolo Cremonesi, "Cross-Domain Recommender Systems", in Association for Computing Machinery (ACM) Conference on Recommender Systems Foster City, Silicon Valley, California, USA, 6th-10th October 2014.

[7] Sheng Gao, Hao Luo, Da Chen, Shantao Li, and Patrick Gallinari, "A Cross-Domain Recommendation Model for Cyber-Physical Systems", in IEEE Transactions Emerging Topics in Computing, vol 1, no. 2, pp. 384-393, December 2013.

[8] Daqiang Zhang, Ching-hsien Hsu, Min Chen, Quan Chen, Naixue Xiong, and Jaime lloret, "Cold-Start Recommendation using Bi-Clustering and Fusion for Large-Scale Social Recommender Systems", in IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 2, pp. 239-250, June. 2014.

[9] M. Sunitha Reddy, and Dr. T. Adilakshmi, "Music Recommendation System based on Matrix Factorization technique –SVD", in International Conference on Computer Communication and Informatics (ICCCI -2014), Jan. 2014.

[10] Fernandez-Tobias, and Cantador, "Exploiting Social Tags in Matrix Factorization Models for Cross-domain Collaborative Filtering", in the First International Workshop on New Trends in Content-based Recommender Systems, pp. 334-347, 2013.

[11] Dimitrios Bouzas, Nikolaos Arvanitopoulos, and Anastasios Tefas, "Graph Embedded Nonparametric Mutual Information For Supervised Dimensionality Reduction", in IEEE Transactions on Neural Networks and Learning Systems, vol. 26, no. 5, pp. 951-963, May. 2015.

[12] Mehmet Gönen, "Bayesian Supervised Dimensionality Reduction", in IEEE Transactions on Cybernetics, vol. 43, no. 6, pp. 2179-2189, Dec. 2013.

[13] Christos Boutsidis, Anastasios Zouzias, Michael W. Mahoney, and Petros Drineas, "Randomized Dimensionality Reduction for *k*-Means Clustering", in IEEE Transactions on Information Theory, vol. 61, no. 2, pp. 1045-1062, Feb. 2015.

[14] Meng Jiang, Peng Cui, Xumin Chen, and Fei Wang, "Social Recommendation with Cross-Domain Transferable Knowledge", in IEEE Transactions on Knowledge and Data Engineering, pp. 1041-4347, 2015.

[15] Bin Li, Xingquan Zhu, Ruijiang Li, and Chengqi Zhang, "Rating Knowledge Sharing in Cross-Domain Collaborative Filtering", in IEEE Transactions on Cybernetics, vol. 45, no. 5, pp. 1054-1068, May 2015.

[16] Fei Hao, Shuai Li, Geyong Min, Hee-Cheol Kim and Stephen S. Yau, "An Efficient Approach to Generating Location-Sensitive Recommendations in Ad-hoc Social Network Environments", in IEEE Transactions on Services Computing, vol. 8, no. 3, pp. 520-533, 2015.

[17] "TripAdvisor: Best Chennai Restaurants", 2016. [Online]. Available:https://www.tripadvisor.in/Restaurants-g304556-Chennai_Madras_Tamil_Nadu.html. [Accessed: 15-Feb-2016].

IRJECE