Multi-dimension Tensor Factorization Collaborative Filtering Recommendation for Academic Profiles

Jesús Silva[1], Noel Varela[2], Omar Bonerge Pineda Lezama[3], Hugo Hernández-P[4], Jairo Martínez Ventura[5], Boris de la Hoz[6], and Leidy Pérez Coronel – Hernández[7]

[1]Universidad Peruana de Ciencias Aplicadas, Lima, Perú.
jesussilvaUPC@gmail.com
[2]Universidad de la Costa, St. 58 #66, Barranquilla, Atlántico, Colombia
nvarela2@cuc.edu.co
[3] Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras
omarpineda@unitec.edu
[4,5,6,7] Corporación Universitaria Latinoamericana, Barranquilla, Colombia
hugohernandezpalma@gmail.com; {academico, bdelahoz }@ul.edu.co;
leidyperezcoronel89@gmail.com

Abstract. The choice of academic itineraries and/or optional subjects to attend is not usually an easy decision since, in most cases, students lack the information, maturity, and knowledge required to make right decisions. This paper evaluates the support of Collaborative Systems for helping and guiding students in this decision-making process, considering the behavior and impact of these systems on the use of data different from the formal information the students usually use. For this purpose, the research applied the clustering based Multi-Dimension Tensor Factorization approach to build a recommendation system and confirm that the increment in tensors improves the recommendation accuracy. As a result, this approach permits the user to take advantage of the contextual information to reduce the sparsity issue and increase the recommendation accuracy.

Keywords: Collaborative filtering, context aware recommendation system, contextual modeling, item recommendations, multi-dimensionality, tensor factorization.

1. Introduction

People are continually making important decisions, sometimes facing many alternatives to consider. There are three main elements that play a fundamental role in the decision-making process: (i) the maturity degree of the individual, (ii) the level of knowledge, and (iii) the information available to make the best decision [1], [2]. Sometimes, inexperienced individuals in a specific field of education may not reach the desirable level of knowledge for making the best choices, so it is important to provide tools to assist them either by providing relevant information or by defining the different options to get orientation for a better decision making.

During the educational training stage of every individual, there are moments when the student must make certain decisions regarding the future. Some questions arise:

what kind of training suits me? what area to choose? what academic itinerary to follow? which subjects to choose ...? This fact is inevitable and happens in most educational stages, starting with Secondary Education in which the degree of responsibility, maturity, and knowledge of the students when making these important decisions is questionable. Is there any way to help students in these proposed tasks either by defining the spectrum of possibilities or by orienting towards an educational itinerary? [3], [4] and [5]. This research intends to answer to these questions by proposing a Recommendation System based on Collaborative Filtering algorithms (hereinafter CF).

A generic multi-dimensional framework based on Tensor Factorization is presented to address context aware recommendations with MD-TFCF (Multi-Dimension Tensor Factorization Collaborative Filtering). Tensor Factorization is used as it can handle any number of contextual variables. Tensor Factorization allows flexible assimilation of contextual information by modeling the context associated with user and products. The contextual information is related to additional dimensions that are represent in the form of tensors. The factorization of this tensor helps in building a unified model of data which provides context aware recommendations. The proposed approach allows integrating more than one context at a time and helps predicting the missing ratings.

The contribution of the research are the following: (1) an efficient 5-mode Tensor Factorization approach is proposed to factorize the tensors, (2) uses Tensor Factorization for the explicit generation of recommendations in which model based clustering and Tensor Factorization learning method is combined to predict missing ratings, (3) Comparative analysis of higher order tensors with lower order tensors is done and confirms that the proposed approach, so the MD-TFCF, leads to more promising results when more contextual dimensions are considered. The results confirm that as the number of contextual dimensions increases, more accurate the recommendations are.

2. Theoretical Review

Various recommendation systems are used on the basis of content based collaborative filtering or hybrid-based approach. Most of the work on CF has been done on traditional 2D-matrix, i.e. user-item rating matrix, but recently, context has become an important factor to be integrated into the recommendation generation algorithms as context plays an important role in real applications such as temporal effect while doing online shopping or selecting places [6]. So, the relevant work of the study in this domain focuses in this point.

Recommendation Systems have been initially devised to improve the decision strategy of users under complex information environments [7] and [8]. Recommendation Systems reduce the problem of information overload by recommending the users most relevant information. Recommendation Systems use content based [9] and [10], collaborative filtering [11], and hybrid filtering [12] techniques for efficient recommendations. The collaborative filtering approach is the most prevailing approach which is further divided into implicit feedback and explicit feedback [13] and [14] methods. In the implicit feedback method, the user's interaction is analyzed in clicks, time spent, and other indicators, and in explicit feedback about the ratings assigned to specific items, questionnaires filled by the user, and others are considered. Then, based on these factors, recommendations are given [15]. CF approach can also be broadly categorized in two types: memory-based, and model-based [16]. In the memory-based method, user or product rating vectors are used to compute analogy among users or products which further operate on a neighborhood-based method. But the major challenge faced in memory-based collaborative filtering approach is the sparsity of the user-item rating matrix, i.e. several entries in the rating matrix might be NULL as there are many non-rated products available in the data pool. This sparsity problem can be reduced by using the model-based approach. In this approach, the generalized model is built to discover latent factors or use the contextual information of users or items for capturing user´s preferences. The most common model-based approach is the Matrix Factorization technique as it considers latent factors that reduce the sparsity of the matrix and gives better results than the User-based Collaborative Filtering approach which simply uses neighborhood approach to find similar users [8]. But the Matrix Factorization technique [16] cannot integrate the contextual information in a straightforward way, so this concept has been extended to multi-dimensional matrices known as Tensor Factorization [15]. that, in this contextual information, can be integrated in more easy ways to give more accurate results than the Matrix Factorization.

The more related work in this domain is elaborated like there are various Tensor Factorization models available which can be used to incorporate contextual information which increases the flexibility and quality of the recommendation systems [17]. Tensor factorization models are applicable in almost every domain due to the increase of computational complexity and the need of a dynamic environment. [18] issued a thorough survey on tensor models, their application domains and the available software. The authors [18] propose various tensor decomposition models such as PARAFAC, DEDICOM, PARATUCK2. Other successful recommendation approaches are the Context-Aware filtering techniques which are broadly categorized as Contextual Pre-Filtering, Contextual Post-Filtering, and Contextual Modeling [15]. The comparative analysis of the three approaches is done by [19] to determine which approach is better and under what situation in relation to accuracy and diversity. The factors considered for evaluating the performance are the dataset type, type of recommendation, and context granularity.

Similarly, [20] presented the Tensor Factorization and Tag Clustering Model (TCM) for recommendations in social tagging systems in which content information is processed to find tags among comparable items, then the tag clusters are formed and finally, association among users, items, and topics are discovered by working upon the Tensor Factorization technique, i.e. Higher Order Singular Value Decomposition (HOSVD). But this work is limited to just three dimensions whereas the proposed approach extends to 5 dimensions and confirms that higher dimensions gives better results. In the same way, [21] proposed a new model Multiverse Recommendation in which contextual information has been integrated with the traditional user-item rating matrix which is not as easy for integrating the contextual information in other model-based approaches like Matrix Factorization. This contextual information represents additional dimensions to original user-item rating matrix as tensor. This approach outperforms other traditional methods which do not involve contextual information in terms of Mean Absolute Error up to 30% whereas the proposed work implements up to

5 dimensions while the performance of proposed recommendation system is assessed against various evaluation metrics.

Recently, [22] introduced the Contextual Modeling Probabilistic Tensor Factorization (CMPTF) model which is basically abstraction of the Probabilistic Tensor Factorization (PTF). In PTF model, the entire information like ratings, item content, context, and social relationship is integrated into a single model which was not possible in earlier approaches. CMPTF further integrates topic modeling information which improves the quality of recommendation systems, and experimental results prove that this approach is superior than traditional approaches. [23] proposed other generic context-aware implicit feedback recommendation algorithms and employ a fast, ALS-based tensor factorization learning method that linearly scales with the number of non-zero elements in the tensor while maintaining the computational efficiency.

Thus, considering the mentioned confrontations by various researchers, the proposed MD- TFCF approach integrates the contextual information as higher order tensors and results support that increment in tensors improves the recommendation performance.

3. Data and Methods

The formal teaching that allows some degree of choice present the following structural patterns: (i) there are students who are enrolled in subjects and obtain certain qualifications; (ii) the subjects are associated to a course, level, or degree, and can be of different types depending on whether they are mandatory, optional, referring to a specific modality or profile, with groupings of subjects that form profiles or educational itineraries in the case of attending to all or a group of them. An academic record can be defined as a set of grades obtained by a student in a series of subjects taken over a certain time period.

The main objective of this contribution is to answer to the following question: is it possible to use people's academic records to offer suggestions when choosing their future? Initially, the answer is not entirely clear since subjective, psychological, and aptitude factors come into play.

Since qualifications provide reliable information about the skills of a student, the areas where people perform best, and even their preferences, a Collaborative Recommendation System is evaluated, estimating the possible qualification that a student would obtain in a subject in case of studying it, to observe if it provides relevant information which, properly linked to future information, could help individuals to make decisions about their future. To this purpose, a series of experiments was conducted to obtain a reliable output to this issue.

3.1 Data

The used data set consists of a total of 7315 anonymous students from primary, secondary, and university levels from several private education institutions in Colombia, considering up to 100 subjects and a total of 155,022 qualifications, which involve values from 0 and 5.

3.2 The Proposed MD-TFCF Mechanism

This section presents the framework of the Multi-Dimension Tensor Factorization Collaborative Filtering (MD-TFCF) approach. The work flow of the proposed framework is shown in Fig. 1 [24], which illustrates that the process starts from the data processing and continues to predictions according to the wishes of the users.

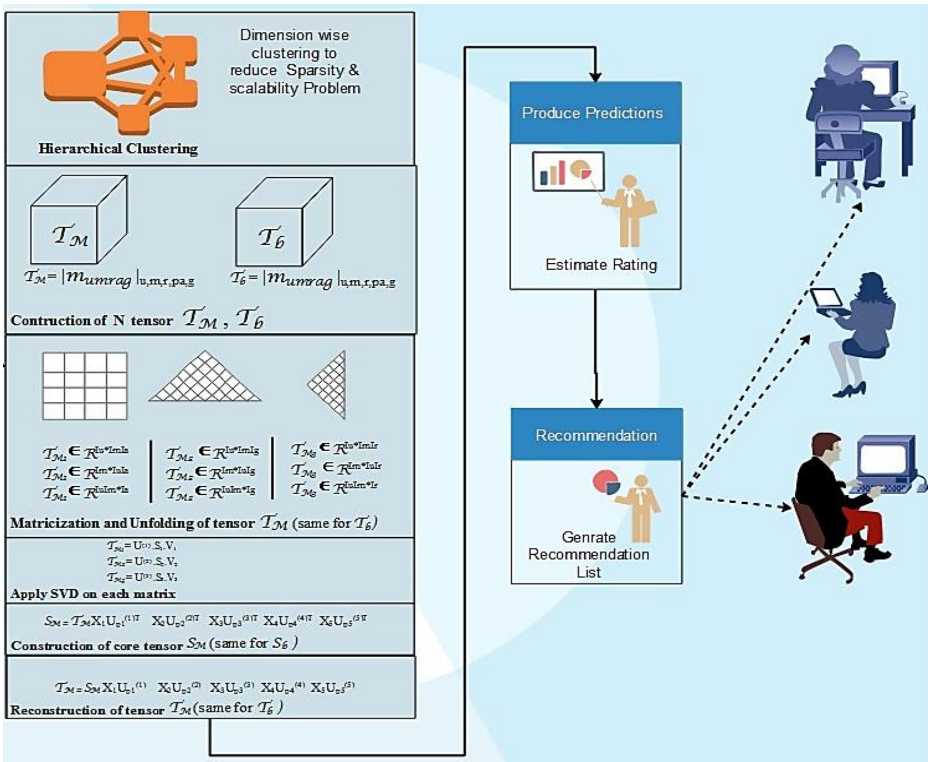

Fig 1. Proposed Multi-Dimension Tensor Factorization Collaborative Filtering (MD-TFCF) Framework, based in Lee, J. et al. (2016) [24]

3.2.1 Hierarchical Clustering Approach

Hierarchical Clustering is one of the coherent clustering techniques [12] in which hierarchies of clusters are formed and every formed cluster is part of another cluster.

This research applies the agglomerative hierarchical clustering-based approach in which the clustering process starts with one initial cluster and then a pair of clusters are merged up together. So, clusters based on age are formed first including users grouped by age, as shown in Table 1 [4].

3.2.2 Decomposition of the singular value of higher order.

In the consulted literature, several tensor decomposition models are available [18] such as PARAFAC, Tucker, Canonical, HOSVD, etc. In the study, the Higher Order

Singular Value Decomposition (HOSVD) Model is used to factor the tensors in matrices obtained from the qualification matrix. The main benefit of using HOSVD is to address the high dimensionality of the data in an effective way [14] and [20], which helps to discover the relationship between users, the qualifications, and other contextual dimensions such as age, gender, and academic term.

Table 1. Categorization according to age

Age Group	Group Name
0-12	Kids
13-17	Teenager
18-25	Youth
26-50	Middle
51-73	Aged

The Higher Order Singular Value Decomposition (HOSVD) Model is constituted by the following stages, for more details see Lee, J. et al. (2016) [24]:

- Initial Construction of Tensor
- Matricization of Tensor (T_i)
- Apply SVD on each matrix (TM_i)
- Construction of Core Tensor (SM)
- Reconstruction of Tensor (TM')
- Recommendation List

3.3 Experimental Setup

The Pareto Principle which is also known as 80/20 rule is used for the verification of the predicted rating allotted through the projected MD-TFCF approach. According to the Pareto Principle the dataset is divided and evenly distributed into training and test set in the ratio of 80% and 20% respectively. The data is evenly distributed in 80-20 ratio so that the entire dimensions data are distributed conceptually. The approach is experimented and assessed on cluster sets formed through the hierarchical clustering approach, for dataset each experiment is run 26 times. Henceforth, the prediction error is minimized using Pareto Principle as it arbitrates in evaluating the efficiency of the proposed MD-TFCF approach.

3.3.1 Evaluation Metrics

The peculiarity of a recommendation algorithm can be assessed using different forms of metrics. The suitability of the metrics used reckons on the recommendation approach, dataset, and what the recommender system will perform. Moreover, Mean Absolute Error (MAE), precision, and recall [13] and [17] are statistical measures to assess the accuracy and peculiarity of the recommendation system.

Mean Absolute Error (MAE): the MAE is the most popular and simplest form of metrics [15] for measuring the accuracy. The MAE basically measures the average absolute difference between the predicted and the actual rating. It is simply, as the name suggests, the mean of the absolute error. It is a measure of deviation of the recommendation or absolute error between the predicted value and the user specific rating value. It is formally calculated using equation (1) as:

$$MAE = \frac{1}{N}\sum_{u,i\in N}\left|p_{u,i} - r_{u,i}\right| \quad \textbf{(1)}$$

Where $p_{u,i}$ is the predicted rating for user u on subject i, $r_{u,i}$ is the actual rating, and N is the total number of ratings. The lower is the value of MAE, the more accurate the recommendation system is for predicting ratings of users. It tells how big an error can be expected from the approach. Other metric measures used for evaluation are classic measure-precision and the recall.

Precision: The Precision is basically the measurement of the probability that the retrieved record is a relevant record [15]. The precision rate is the fraction of successful rating prediction that is predicted by users. The precision is computed using the equation (2) as:

$$precision = \frac{Correctly\ Predicted\ Rating}{Total\ No.of\ Correctly + Incorrectly\ Predicted\ Rating} \quad \textbf{(2)}$$

Therefore, the precision identifies the ratio of the number of the correctly predicted rating retrieved to the total number of incorrectly and correctly predicted ratings.

Recall: It is defined as fraction of relevant prediction retrieved to the total number of the user prediction in the dataset. The recall is computed using the equation (3) as:

$$Recall = \frac{Correctly\ Predicted\ Rating}{Total\ number\ of\ User\ Assigned\ Prediction} \quad \textbf{(3)}$$

4. Results and Discussions

The proposed MD-TFCF approach is different from existing approaches as an integrated framework is developed in the proposed approach to unanimously represent the five dimensions. Figure 2 shows that there are remarkable improvements in results in form of precision, recall, and mean absolute error for the datasets.

Figure 2 infers that precision varies from 0.54 to 0.96; recall varies from 0.30 to 0.80, and the mean absolute error decreases from 2.2 to 0.38 for dataset, while similarly, precision varies from 0.753 to 0.916, recall varies from 0.50 to 0.73, and the mean absolute error decreases from 2.2 to 0.38 showing that the MD-TFCF approach achieves more promising results than the traditional user-item based collaborative filtering approach. In the same way, on adding even one dimension, i.e. 5-tensor approach, is better than 4-tensor as accuracy in results has been improved as precision varies from 0.50 to 0.77, recall varies from 0.30 to 0.60, and the mean absolute error decreases from 1.86 to 1.02. Thus, a new technique is concurrently proposed to deal with 5 dimensions and used for comparative analysis with traditional user-item based approach and with lower dimensional spaces.

Measuring Metrics →	Recall			Precision			MAE		
Approaches ↓	Max	Min	Avg	Max	Min	Average	Max	Min	Avg
User-Item Based Neighborhood Collaborative Filtering									
Normal U-I	0.5	0.1	0.23	0.5	0.05	0.38	3	1.2	2.4
Multi-Dimensional Tensor Factorization Collaborative Filtering									
4-order Tensor	0.6	0.30	0.50	0.77	0.50	0.65	1.86	1.02	1.46
5- order Tensor	0.80	0.30	0.65	0.96	0.54	0.81	2.2	0.38	1.06

Fig 2. Comparative Analysis of Higher Order Tensor with Lower Order Tensor Results

It is empirically validated that MD-TFCF approach gains about 49% accuracy in form of precision, 20% in form of recall and 32% in terms of mean absolute error for the studied dataset. Thus, the proposed approach is achieving more desirable results whenever more contextual parameters are considered. Figure 3 shows results of conventional user-item based neighborhood CF process and MD-TFCF (higher order tensors with lower order tensors) approach in comparison to each other in form of graph. As shown in Fig. 3, conventional algorithm's precision and recall varies from 5% - 50% and 1% - 5% respectively, for dataset.

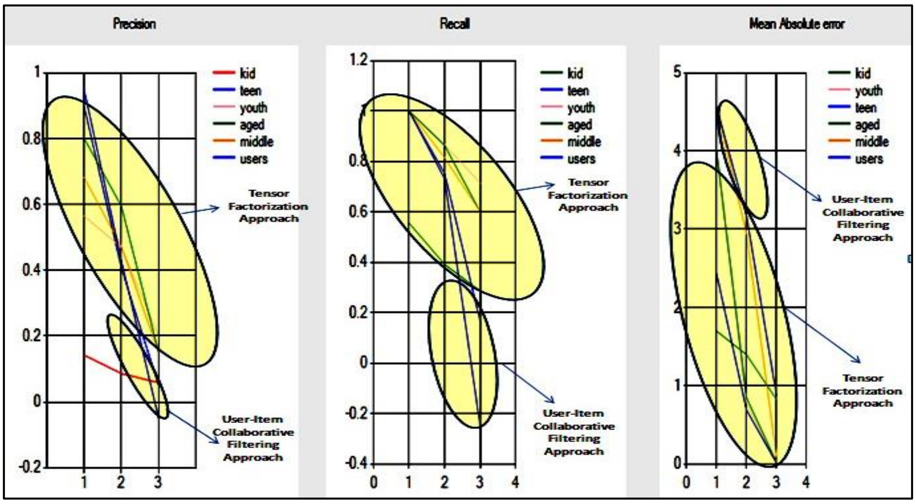

Fig 3. Comparative Analysis Results for the set of data studied

The precision and recall values show improvement in dataset because of the large data size. The following graphs (x-axis represent number of folds and y-axis represents recall, precision and mean absolute error respectively) validate that tensor factorization approach provides more accuracy in results in form of precision, recall, and mean absolute error as evaluation metrics than traditional.

5. Conclusions

In this research, a novel Multi-Dimension Tensor Factorization Collaborative Filtering (MD-TFCF) approach is introduced to mitigate the sparsity problem as this is the major challenge of the Collaborative Filtering approach. In traditional user-item based Collaborative Filtering approach, the user-item matrix is formed by considering only ratings accredited by users to different products, but several entries in rating matrix are NULL because there are diverse set of items that are generally not rated by users. So, to overcome this problem, User-Item based approach is extended to Model based approach MD-TFCF and mainly comparative analysis of MD-TFCF with user-item based collaborative filtering and lower order dimensional spaces is done.

After analyzing the recommendation systems based on proposed collaborative filtering, it has been proved that their use can be useful for making personalized recommendations to students about educational itineraries [25] when choosing optional subjects and foreseeing which common subjects will present greater learning difficulties or specific needs of reinforcement in the student.

References

1. Vásquez, C., Torres, M., Viloria, A.: Public policies in science and technology in Latin American countries with universities in the top 100 of web ranking. J. Eng. Appl. Sci **12**(11), 2963–2965 (2017).
2. Aguado-López, E., Rogel-Salazar, R., Becerril-García, A., Baca-Zapata, G.: Presencia de universidades en la Red: La brecha digital entre Estados Unidos y el resto del mundo. Revista de Universidad y Sociedad del Conocimento **6**(1), 1–17 (2009).
3. Torres-Samuel M., Vásquez C., Viloria A., Lis-Gutiérrez, J.P., Borrero, T.C., Varela, N.: Web Visibility Profiles of Top100 Latin American Universities. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, Springer, Cham **10943** (1), 1-12 (2018).
4. Viloria A., Lis-Gutiérrez, J.P., Gaitán-Angulo, M., Godoy, A.R.M., Moreno, G.C., Kamatkar, S.J.: Methodology for the Design of a Student Pattern Recognition Tool to Facilitate the Teaching – Learning Process Through Knowledge Data Discovery (Big Data). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, Springer Cham **10943** (1), 1-12 (2018).
5. Caicedo, E.J.C., Guerrero, S., López, D.: Propuesta para la construcción de un índice socioeconómico para los estudiantes que presentan las pruebas Saber Pro. Comunicaciones en Estadística **9** (1), 93–106 (2016).
6. Reihanian, A., Minaei-Bidgoli, B., Alizadeh, H.: Topic-oriented community detection of rating-based social networks. Journal of King Saud University-Computer and Information Sciences **28**(3), 303-310 (2016).

7. Rashid, A. M., Albert, I., Cosley, D., Lam, S. K., McNee, S. M., Konstan, J. A., Riedl, J.: Getting to know you: learning new user preferences in recommender systems. In Proceedings of the 7th international conference o Intelligent user interfaces. ACM. 127-134 (2002).
8. Gómez, S., Zervas, P., Sampson, D. G., Fabregat, R.: Context-aware adaptive and personalized mobile learning delivery supported by UoLmP. Journal of King Saud University-Computer and Information Sciences **26**(1), 47-61 (2014).
9. Lops, P., De Gemmis, M., Semeraro, G.: Content-based recommender systems: State of the art and trends. In Recommender systems handbook Springer US (2011).
10. Malik, F., Baharudin, B.: Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the DCT domain. Journal of king saud university-computer and information sciences **25**(2), 207-218 (2013).
11. Su, X., Khoshgoftaar, T. M.: A survey of collaborative filtering techniques. Advances in artificial intelligence 4 (1), 1-12 (2009).
12. De Campos, L. M., Fernández-Luna, J. M., Huete, J. F., Rueda-Morales, M. A.: Combining content-based and collaborative recommendations: A hybrid approach based on Bayesian networks. International Journal of Approximate Reasoning **51**(7), 785-799 (2010).
13. Isinkaye, F. O., Folajimi, Y. O., Ojokoh, B. A.: Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal **16**(3), 261-273 (2015).
14. Buder, J., Schwind, C.: Learning with personalized recommender systems: A psychological view. Computers in Human Behavior **28**(1), 207-216 (2012).
15. Taneja, A., Arora, A.: Cross domain recommendation using multidimensional tensor factorization. Expert Systems with Applications **92** (1), 304-316 (2018).
16. Buder, J., Schwind, C.: Learning with personalized recommender systems: A psychological view. Computers in Human Behavior **28**(1), 207-216 (2012).
17. Frolov, E., Oseledets, I.: Tensor methods and recommender systems. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **7**(3), 1-12 (2017).
18. Kolda, T. G., Bader, B. W.: Tensor decompositions and applications. SIAM review **51**(3), 455-500 (2009).
19. Panniello, U., Tuzhilin, A., Gorgoglione, M.: Comparing context-aware recommender systems in terms of accuracy and diversity. User Modeling and User-Adapted Interaction **24**(2), 35-65 (2014).
20. Rafailidis, D., Kefalas, P., Manolopoulos, Y.: Preference dynamics with multimodal user-item interactions in social media recommendation. Expert Systems with Applications **74** (1), 11-18 (2017).
21. Karatzoglou, A., Amatriain, X., Baltrunas, L.,Oliver, N.: Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In Proceedings of the fourth ACM conference on Recommender systems ACM. 79-86. (2010)
22. Zheng, C., Haihong, E., Song, M., Song, J.: CMPTF: Contextual Modeling Probabilistic Tensor Factorization for recommender systems. Neurocomputing **205** (1), 141-151 (2016).
23. Hidasi, B., Tikk, D.: Fast ALS-based tensor factorization for context-aware recommendation from implicit feedback. Machine Learning and Knowledge Discovery in Databases, 67-82 (2012).
24. Lee, J., Lee, D., Lee, Y. C., Hwang, W. S., Kim, S. W.: Improving the accuracy of top-n recommendation using a preference model. Information Sciences **348** (1), 290-304 (2016).
25. Vásquez, C., Torres-Samuel, M., Viloria, A., Lis-Gutiérrez, J.P., Crissien, B.T., Varela, N., Cabrera, D.: Cluster of the Latin American Universities Top100 According to Webometrics 2017. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, Springer, Cham **10943** (1), 1-12 (2018).