Recommender System Based on Latent Topics

María Emilia Charnelli[1,3](✉), Laura Lanzarini[2], Javier Díaz[1]

[1]LINTI - Research Laboratory in New Information Technologies
[2]III LIDI - Computer Science Research Institute LIDI
Computer Science School, National University of La Plata
[3] CONICET - National Scientific and Technical Research Council
mcharnelli@linti.unlp.edu.ar, laural@lidi.info.unlp.edu.ar,
jdiaz@unlp.edu.ar

Abstract. Collaborative filtering is one of the most used techniques in recommender systems. The goal of this paper is to propose a new method that uses latent topics to model the items to be recommended. In this way, the ability to establish a similarity between these elements is incorporated, improving the performance of the recommendation made. The performance of the proposed method has been measured in two very different contexts, yielding satisfactory results. Finally, the conclusions and some future lines of work are included.

Keywords: Recommender Systems, Collaborative Filtering, Latent Topic Modeling.

1 Introduction

Recommender systems analyze patterns of interest to users such as articles or products, to provide personalized recommendations that satisfy their preferences [1]. Suggestions intervene in various decision-making processes, such as which items to buy, which movies to watch, or which books to read. The term item is used to indicate what the system recommends to users [2]. For this purpose, it is necessary to model the items that are to be recommended. Generating of a model from the textual and unstructured information of a set of items represents a great challenge. The analysis of latent topics has emerged as one of the most efficient methods for classifying, grouping and retrieving textual data. Discovering topics in short texts is crucial for a wide range of tasks that analyze topics, such as characterizing content, modeling user interest profiles, and detecting latent or emerging topics. The BTM biterm topic model [3] allows to efficiently extract the topics that characterize a set of short texts. BTM can obtain the underlying topics in a set of documents and a global distribution of each topic within each of them, through the analysis of the generation of biterms.

The most common approach for a recommender system is the collaborative filtering technique based on neighborhood models. Its original form is based on the similarities between users [4]. These user-user methods estimate unknown scores based on registered scores of like-minded users. Subsequently, the analogous approach became popular but now taking into account the similarities between items [5] [6]. In these methods, a score is calculated using assessments made by the same user on similar items. Better scalability and improved accuracy make the item approach more favorable in many cases [7] [8]. In addition, item-item methods are more likely to explain the reasoning behind the predictions. This is because the users are familiar with the elements previously preferred by them, but do not know the supposedly similar users. Most item-item approaches use a measure of similarity between the ratings they have.

This paper proposes a method based on the item-item approach that uses a model of latent topics to model the items that need to be recommended and establishes a similarity between these elements that improve the performance of the recommendation. The evaluation of the proposed method is done through a set of educational materials from the Merlot [9] digital repository and a movie dataset from MovieLens [10]. This article is organized as follows: the second section describes the extraction and modeling of latent topics, the third section describes the proposed method, the fourth section shows the experimental results. Finally, the fifth section presents the conclusions and future lines of work.

2 Extraction of Latent Topics

For the extraction of the topics in the descriptions of the items, BTM (Biterm Topic Model) was used, which is an unsupervised learning technique that discovers the topics that characterize a set of brief documents.

Let a set of N_D documents be called corpus where W is a set of all the words of the corpus, a topic is defined as a probability distribution over W. Therefore, a topic can be characterized by its T most likely words. Given a K number of topics, the objective of BTM is to obtain the K distributions on each of the words. A "biterm" denotes a pair of words without order that co-occur in a short document. In this case, two different words in a document construct a biterm.

Given a corpus with N_D documents and a W unique-word vocabulary, it is assumed to contain N_B biterms $\boldsymbol{B} = \{b_i\}_{i=1}^{N_B}$ with $b_i = (w_{i,1} \in W, w_{i,2} \in W)$, and K topics expressed over W. Let $z \in [1, K]$ be a variable to indicate a topic. The probability $P(z)$ that a document in the corpus is of a topic z, defined as a multinomial K-dimensional distribution $\boldsymbol{\theta} = \{\theta_k\}_{k=1}^{K}$ with $\theta_k = P(z = k)$ and $\sum\limits_{k=1}^{K} \theta_k = 1$. The distribution of words by topic $P(w|z)$ can be represented as a matrix $\varPhi \in R^{K \times W}$ where the kth row ϕ_k is a multinomial distribution W-dimensional with input $\phi_{k,w} = P(w|z = k)$ and $\sum\limits_{w=1}^{W} \boldsymbol{\phi}_{k,w} = 1$. Given the parameters α and β, the main assumption of the model is that each of the documents of the corpus were generated in the following way:

1. A topic distribution $\boldsymbol{\theta} \sim Dirichlet(\alpha)$ is chosen for all the corpus
2. For each topic $k \in [1, K]$
   - A word distribution is extracted for the topic $\phi_k \sim Dirichlet(\beta)$
3. For each biterm $b_i \in \boldsymbol{B}$

- A topic assignment is extracted $z_i \sim Multimonial(\boldsymbol{\theta})$
- Two words are extracted $w_{i,1}, w_{i,2} \sim Multimonial(\phi_{z_i})$

Taking into account the generation mechanism assumed by BTM, the likelihood can be obtained for the entire corpus given the parameters α and β from the probability of each of the biterms:

$$P(\boldsymbol{B}|\alpha,\beta) = \prod_{i=1}^{N_B} \int\int \sum_{k=1}^{K} P(w_{i,1}, w_{i,2}, z_i = k|\boldsymbol{\theta}, \boldsymbol{\Phi}) d\boldsymbol{\theta} d\boldsymbol{\Phi} \tag{1}$$

$$= \prod_{i=1}^{N_B} \int\int \sum_{k=1}^{K} \theta_k \phi_{k,w_{i,1}} \phi_{k,w_{i,2}} d\boldsymbol{\theta} d\boldsymbol{\Phi} \tag{2}$$

Obtaining exactly the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$ that maximize the likelihood of (2) is an intractable problem. Following the proposal in [11], the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$ can be approximated using Gibbs sampling [12].

To infer the topics of a document, that is, to evaluate $P(z|d)$ for the document d, the proportion of topics of a document is derived through the topics of the biterms. If d contains N_d biterms, $\{b_i^{(d)}\}_{i=1}^{N_d}$,

$$P(z|d) = \sum_{i=1}^{N_d} P(z|b_i^{(d)}) P(b_i^{(d)}|d) \tag{3}$$

2.1 Evaluation Criterion

To evaluate the quality of the topics obtained, the coherence metric proposed by Mimno et al. [13] is used. Given a topic z and its T most probable words $V^{(z)} = (v_1^{(z)}, ..., v_T^{(z)})$ where $v_i^{(z)} \in W$ for $i = 1...T$, the coherence score is defined as:

$$C(z; V^{(z)}) = \sum_{t=2}^{T} \sum_{l=1}^{t} \log \frac{D({v_t}^{(z)}, {v_l}^{(z)}) + 1}{D({v_l}^{(z)})}$$

where $D(v)$ is the frequency of the word v in all documents, $D(v, v')$ is the number of documents where the words v and v' co-occur. The coherence metric is based on the idea that words that belong to the same concept will tend to co-occur within the same documents. This is empirically demonstrable because the coherence score is highly correlated with the human criterion.

To evaluate the overall quality of a set of topics, the average of the coherence metric is calculated for each of the topics obtained $\frac{1}{k} \sum_k C(z_k; V^{(z_k)})$. These results allow us to determine the number of topics that best represent the entire corpus.

3 Method Proposed

Let K be the number of topics that represent a set of items, each of them is modeled according to the probability distribution shown in (3).

Given a list of m users $U = u_1, u_2, .., u_m$ and a list of n items $I = i_1, i_2, ..., i_n$, each user has a list of items I_u, with a score associated to each item r_{ui}. Each item is assigned a score of 1 to 5.

In order to evaluate the similarity between two items from the probability distributions obtained with BTM, the proposed method uses the divergence of Kullback-Leibler [14]. Given two probability distributions P and Q, the divergence function is defined as:

$$D_{KL}(P, Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

From this divergence, it is possible to define the similarity between two items p and q as follows:

$$sim(p, q) = \exp(-D_{KL}(p, q))$$

To estimate the rating of a new item m given a user u, the following method is proposed to predict $\hat{r}_{um}$:

1. The probability distributions are obtained for each of the items that the user evaluated I_u, as shown in the 2 section.
2. The probability distribution of the material m is obtained.
3. The similarity sim of m is calculated with each I_{u_j}.
4. The similarities are ordered, and the first t are chosen, where t is a parameter that defines the size of the neighborhood to be considered.
5. From the most similar t, the prediction is calculated:

$$\hat{r}_{um} = \mu_m + \frac{\sum\limits_{j=1}^{t} sim(m, j)(r_{uj} - \mu_j)}{\sum\limits_{j=1}^{t} sim(m, j)}$$

where r_{uj} is the score of item j given by user u, and μ_j and μ_m are the average scores of j and m respectively.

4 Experimental Results

Two databases were used in this work, one of educational materials and the other one of films. The first one provides information on users and Computer Science educational materials in the Merlot [9] digital repository. The data involves more than 984 materials and more than 260 users who uploaded, evaluated or commented on each of the publications. Also, public information about publications and users was available. The data extracted from each of the publications was: title, type of material, date of creation, date of update, user who made it, reviewer review from 1 to 5, user review from 1 to 5, comments, and the unstructured textual description. The second dataset is about movie ratings in MovieLens[10]. This dataset contains 100,000 scores from 1 to 5 by 943 users for 1682 movies, where each user rated at least 20 movies; of the films the title and the date are known; and in addition, the arguments of each of them were collected.

When it comes to operating with textual information, it is necessary to resort to Text Mining techniques in order to represent each description in a vector of terms. This was carried out through a process consisting of several stages. In a first stage, the contents were unified in a single language. Then a stopwords filter was applied, which is responsible for filtering the words that match any indicated stopword. English language stopwords were filtered; words relating to the context. URLs and non-text characters were also deleted. Then, each word in the text was reduced to its root by applying the Snowball [15] stemming algorithm. The importance of this process is that it eliminates syntactic variations related to gender, number and verbal time. Once the roots of each of the words are obtained, the frequency of appearance of each of them in the publications was calculated and the words that appear more than once were chosen.

From the structured textual information, the modeling of latent topics was obtained through BTM for the set of educational materials and films. To evaluate these models, for each number of topics between 2 and 30, the coherence obtained was averaged, randomly sampling the test and training set in 1000 iterations. Figure 1 shows the average of the coherence of the model with respect to the quantity of topics extracted in the material dataset. The number of topics in which there is a break in the growth of the function of average coherence is of interest. In this case, the optimal value is between 5 and 7 latent topics.

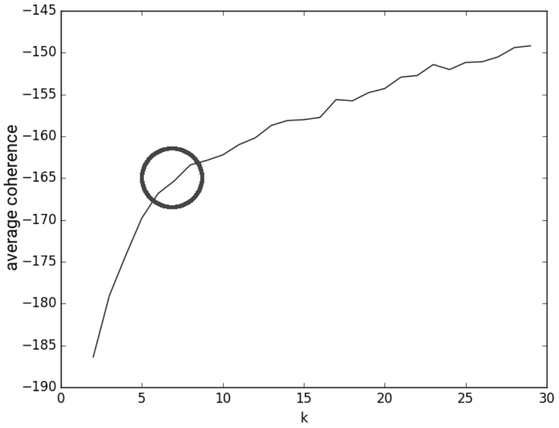

Fig. 1. Materials dataset. Average coherence for different Ks

Table 1 shows the topics obtained with $K = 7$ for the materials dataset. For each of the topics, the six most important words are shown, i.e., those that are more likely to belong to that topic.

Table 1. Materials dataset. Model of topics obtained with BTM

Topic	Most important words in the topic				
1	programming	software	data	algorithms	design
2	information	technology	computing	internet	systems
3	programming	java	language	tutorial	software
4	resources	design	systems	development	security
5	design	information	programming	interaction	human
6	binary	fractions	codes	numbers	tutorial
7	numbers	stars	interactivate	graph	simulation

The evaluation methodology for the proposed method applies 10-fold cross-validation. This evaluation process was repeated 50 times to obtain a significant sample on which the results are averaged. This process was applied to the proposed method, identified as KNN Topic Model, and the KNN collaborative filtering methods, KNN Mean [7] based on the item-item and user-user approach, SlopeOne [16] and on the method based on latent factor models (SVD) [17].

The proposed method and KNN methods receive as a parameter the number of neighbors to consider. The size of the neighborhood has a significant impact on the quality of the prediction [4]. Figure 2 shows the RMSE (Root Mean Squared Error) for different numbers of neighbors in the different algorithms. The error decreases as the number of neighbors grows. The error for the proposed KNN Topic Model method is always below for different neighborhood values. In addition, it is observed that after 40 neighbors the RMSE decreases slowly for each of the algorithms.

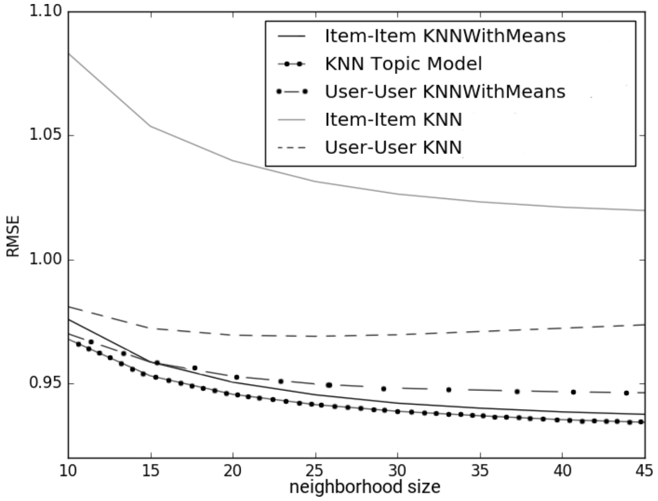

Fig. 2. Movies dataset. Influence of the neighborhood size

The results of the 50 executions of the cross validation for each algorithm using the material dataset are shown in the Table 2 and the results when using the movie dataset are shown in the Table 3. The number of neighbors $t = 40$ was established for all neighborhood-based models. The items of the material dataset were represented as a multinomial 7-dimensional distribution and the items of the movie dataset as a 10-dimensional multinomial distribution. To evaluate the predictions of the proposed method against the results of the other algorithms, the precision metrics calculated were RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) and FCP (Fraction of Concordant Pairs), which measures the proportion of pairs of well-classified items [18]. Unlike RMSE and MAE, the value of FCP is better the higher it is, since it measures a proportion.

Table 2. Educational Materials dataset. Results obtained

	KNN Model Topic	KNN item-item	KNN Mean item-item	KNN user-user	KNN Mean user-user	SlopeOne	SVD
Mean RMSE	**0.6047**	0.6848	0.8412	0.7757	0.6339	0.6575	0.6420
Mean MAE	0.4403	0.4566	0.5544	0.5126	0.4333	0.4336	**0.3443**
Mean FCP	**0.6517**	0.2075	0.4820	0.1400	0.3333	0.4133	0.4329

It is observed that the proposed method is competitive against two sets of different datasets. For the dataset of educational materials, the KNN Topic Model

Table 3. Movies dataset. Results obtained

	KNN	KNN	KNN	KNN	KNN	SlopeOne	SVD
	Model	item-item	Mean	user-user	Mean		
	Topic		item-item		user-user		
Mean RMSE	**0.9340**	1.0203	0.9385	0.9732	0.9466	0.9426	0.9402
Mean MAE	**0.7359**	0.8044	0.7375	0.7680	0.7462	0.7409	0.7396
Mean FCP	0.6879	0.5990	0.6867	0.6948	**0.6946**	0.6865	0.6889

method obtains a lower RMSE error and a higher FCP ratio. However, the MAE metric is lower for SVD. It is emphasized that with the little information of the materials that are available, through the use of topic modeling it is possible to improve the FCP. In the movie dataset, more information about the interests of the users is available, so that the proposed method, although having a competitive result, does not exceed the FCP value with respect to the user-user KNN Mean approach.

5 Conclusions and Future Lines of Work

In the present paper we managed to model a set of items detecting latent topics in their descriptions. This allowed us to know which are the topics that describe the items and how they relate to each other. The methodology used in the proposed method and the validation metrics applied present preliminary results that are satisfactory and competitive compared to traditional methods. As future work, the application of the proposed method in other databases with associated textual information is foreseen. It is also interesting to incorporate information about the opinions and tastes of the user from other contexts. The results of this work are an addition to the work previously presented in [19], where a modeling of users was proposed through the information obtained with BTM when identifying the topics of interest of the students of the Computer Science School of the UNLP through their publications made in Facebook groups. In turn, this work is related to a larger project, whose objective is to create a recommender system for digital educational materials.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE transactions on knowledge and data engineering **17**(6) (2005) 734–749
2. Ricci, F., Rokach, L., Shapira, B.: Introduction to recommender systems handbook. In: Recommender systems handbook. Springer (2011) 1–35
3. Cheng, X., Yan, X., Lan, Y., Guo, J.: Btm: Topic modeling over short texts. IEEE Transactions on Knowledge and Data Engineering **26**(12) (2014) 2928–2941

4. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM (1999) 230–237
5. Linden, G., Smith, B., York, J.: Amazon. com recommendations: Item-to-item collaborative filtering. IEEE Internet computing **7**(1) (2003) 76–80
6. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th international conference on World Wide Web, ACM (2001) 285–295
7. Bell, R.M., Koren, Y.: Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In: Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, IEEE (2007) 43–52
8. Takács, G., Pilászy, I., Németh, B., Tikk, D.: Major components of the gravity recommendation system. ACM SIGKDD Explorations Newsletter **9**(2) (2007) 80–83
9. University, C.S.: Merlot - multimedia educational resource for learning and online teaching. `https://merlot.org` (2017 (accessed June 30, 2017))
10. Researc, G.: Movielens datasets. `https://grouplens.org/datasets/movielens/` (2017 (accessed June 30, 2017))
11. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National academy of Sciences **101**(suppl 1) (2004) 5228–5235
12. Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence (6) (1984) 721–741
13. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics (2011) 262–272
14. Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics **22**(1) (1951) 79–86
15. Gupta, V., Lehal, G.S.: A survey of common stemming techniques and existing stemmers for indian languages. Journal of Emerging Technologies in Web Intelligence **5**(2) (2013) 157–161
16. Lemire, D., Maclachlan, A.: Slope one predictors for online rating-based collaborative filtering. In: Proceedings of the 2005 SIAM International Conference on Data Mining, SIAM (2005) 471–475
17. Mnih, A., Salakhutdinov, R.R.: Probabilistic matrix factorization. In: Advances in neural information processing systems. (2008) 1257–1264
18. Koren, Y., Sill, J.: Collaborative filtering on ordinal user feedback. In: IJCAI. (2013) 3022–3026
19. Charnelli, M.E., Lanzarini, L., Diaz, J.: Modeling students through analysis of social networks topics. XXII Congreso Argentino de Ciencias de la Computacion CACIC 2016 (2016) 363–371