NEWS RECOMMENDATION BASED ON COLLABORATIVE SEMANTIC TOPIC MODELS AND RECOMMENDATION ADJUSTMENT

YU-SHAN LIAO, JUN-YI LU, DUEN-REN LIU

Institute of Information Management, National Chiao Tung University, Hsinchu 300, Taiwan
E-MAIL: allenysliao@gmail.com, alex42132000@hotmail.com, dliu@mail.nctu.edu.tw

Abstract:

Providing news recommendations is an important trend for online news websites to attract more users and create more benefits. In this research, we propose a novel recommendation approach for recommending news articles. We propose A Collaborative Semantic Topic Model and an ensemble model to predict user preferences based on combining Matrix Factorization with articles' semantic latent topics derived from word embedding and Latent Dirichlet Allocation. The proposed ensemble model is further integrated with a recommendation adjustment mechanism to adjust users' online recommendation lists. We evaluate the proposed approach via offline experiments and online evaluation on a real news website. The experimental result demonstrates that our proposed approach can improve the recommendation quality of recommending news articles.

Keywords:

Recommendation; Latent topic analysis; Collaborative topic model; Recommendation adjustment

1. Introduction

With the ubiquity of Internet, more users receive news from the online websites. New types of media platforms provide various topics such as lifestyle and fashion news. Though the online news platform becomes a significant channel for users to retrieve required information, information overloading makes it difficult to fulfill users' demands. Accordingly, recommendation systems are needed to increase the users' loyalty and browsing aspiration. With more network traffic, news platforms may get more revenue.

There are news articles posted on the news website every day. For news recommendation, the recommender system needs to handle the issues of recommending cold start news articles with very few browsing records. Existing researches on news recommendations focused on analyzing users' history data, including their rating or click status to achieve good recommendation result [1-5]. Among these methods, collaborative filtering (CF) [6, 7] such as Matrix Factorization (MF) [8] is a typical way to extract the implicit factors for predicting the preference ratings of users. But MF confronts the cold-start and data sparsity problems. Hence, Collaborative Topic Model (CTM) [9, 10] combines content-based filtering (CBF) like Latent Dirichlet Allocation (LDA) [11] with MF to strengthen the latent factor analysis and tackle the issue of cold-start and data sparsity. Nevertheless, news are not only constituted by content topics, semantic vectors of word embedding play an important role to link the whole text [12]. Combining latent topic distribution with word semantic vectors is more integral for discovering the implicit factors for news articles. Moreover, online websites have limited displaying layout for online recommendation, and thus it is important to dynamically adjust the recommendation lists.

In this work, we propose a novel recommendation approach that considers user preference analysis and online recommendation adjustment (RA). In the user preference analysis, semantic LDA model (SLDA) is used to extract the semantic latent topics of news articles by integrating the semantic vectors of word embedding with LDA. Moreover, Collaborative Semantic Topic Model (CSTM) is proposed to predict user preferences by combining the SLDA and MF. An ensemble model (EM) which combines SLDA-based and CSTM-based prediction models is proposed to predict user preferences.

In online recommendation, we propose an online news recommendation approach by further integrating the proposed ensemble model with recommendation adjustment (RA) mechanism. We adopt the RA mechanism [13] to adjust the recommendation list based on the push and click frequency of target news. Accordingly, our proposed approach can deliver more user-interested news without restricting the number of recommended news because of limited displaying layout. We evaluate our proposed approach not only via the offline experiments, but also through the online experiments in a real website. The offline experiments and online evaluation show that our proposed CSTM and EM methods can improve the recommendation quality and obtain better performance than typical CTM method. The online evaluation also shows that the proposed EM method integrated with recommendation adjustment can

improve the click through rate for online news recommendation.

2. Related work

2.1 Recommender systems

Recommender systems have been broadly applied in recommending articles [3, 5, 10]. The content-based filtering (CBF) focuses on recommending items that are similar to users' interests based on calculating the similarities between user profiles and items [14, 15]. The collaborative filtering (CF) recommends items based on neighbors who have similar preferences by measuring the similarities between users [7]. The latent factor models adopt matrix factorization techniques to discover the latent factor vectors for users and items [8].

2.2 Latent Dirichlet Allocation and word embedding

LDA is a generative probabilistic topic model to analyze the thematic structure and extract potential variables from corpora. In LDA, documents are constituted as random mixtures on latent topics, and each topic is characterized by a distribution of words [11]. Thus, the topic of each document becomes multinomial. To simplify the complex process of LDA model, Gibbs sampling is used to solve the problem of inefficient convergence speed [16]. Mikolov et al. [17] embed the high dimensional vector space into low dimensional continuously vector space to define the representation for each word. Semantic coherence is helpful for chaining the meaning in the whole text. Thus, certain studies conform topic modeling with word embedding. Das et al. [12] apply multiple sampling methods and treat word embedding as an input of LDA to improve the topic coherence of documents based on measuring the pointwise mutual information of topic words.

2.3 Collaborative Topic Model

Wang and Blei [9] present a Collaborative Topic Model (CTM), combining LDA with Matrix Factorization (MF) to improve the recommendation of scientific articles. CTM uses topic vectors extracted by LDA to build the latent factors. CTM is based on MF, which factorizes the user-item rating matrix into two factors: item latent factor and user latent factor [8]. CTM uses LDA to build K dimensional latent topic θ_j for each article. Besides, according to user's article rating matrix, CTM uses MF to build K dimensional latent factor u_i for each user and latent factor v_j for each article by setting $v_j = \theta_j + \varepsilon_j$, where ε_j is the rating latent factor for each article. Multiplying the two latent factors can approximate the original rating matrix, the predicted rating $\hat{r}_{i,j} = u_i^T v_j$, where u_i^T is the transposed user latent factor and v_j denotes the item latent factor.

3. Proposed approach

3.1 Overview

An overview of the proposed recommendation approach is shown in FIGURE 1. News articles are processed to document vectors by TFIDF and word embedding vectors. Then, Semantic LDA (SLDA) is conducted to derive the semantic latent topic vectors of news articles. Next, we analyze user preferences by using SLDA and Collaborative Semantic Topic Model (CSTM) to derive user preferences. An ensemble model combining SLDA and CSTM prediction models is also proposed to calculate user interest scores. In online recommendation, the proposed approach adopts a recommendation adjustment mechanism to adjust the recommendation list.

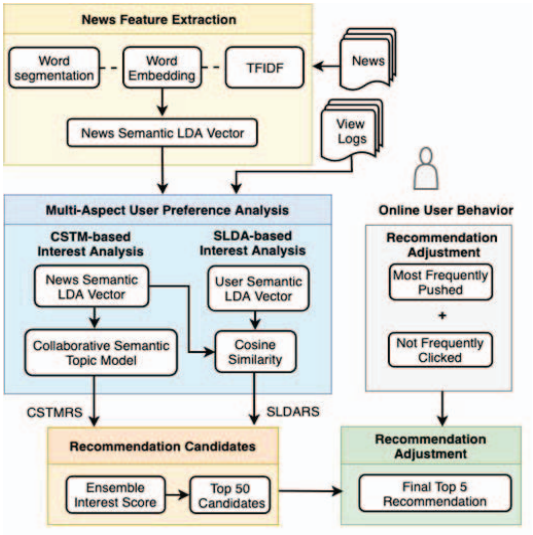


FIGURE 1. An overview of the proposed recommendation method

3.2 News feature extraction

Some terms are more important in representing news articles, so we give terms in news with distinct weights, as listed in Eq. (1).

$$\alpha_{d,i} = \begin{cases} 1, & \text{if term } i \text{ is a content word} \\ 2, & \text{if term } i \text{ is a title or emphasized word} \\ 3, & \text{if term } i \text{ is a tag} \end{cases} \quad (1)$$

Traditional LDA derives the latent topics of news based on TFIDF term vectors of news. We derive the weighted TFIDF term of news d, $wt_{d,i}$, by Eq. (2). $tf_{d,i}$ represents the term frequency of term i in news d and idf_i is the inverse document frequency of term i.

$$wt_{d,i} = \alpha_{d,i} \times tf_{d,i} \times idf_i \quad (2)$$

Semantic LDA (SLDA) is adopted to derive the semantic latent topics of news by utilizing the semantic vectors of word embedding as the input of LDA. Let $\overrightarrow{we_d}$ represent the k dimensional word embedding vector for news d. We employ Gensim word2vec to build the word embedding by utilizing Chinese Wiki documents and NIUSNEWS articles. Two kinds of SLDA models are proposed. The WE method derives the word embedding vector for news d, as the summation of the word embedding $\overrightarrow{e_{d,i}}$ multiplied with the term weight $\alpha_{d,i}$ of term i for $i \in W_d$, the set of terms in news d.

The terms' *tfidf* values may affect the performance of deriving the semantic latent topics. The WE_TFIDF method uses *tfidf*-based word embedding vector for news d, $\overrightarrow{we\_tfidf_d}$ by using non-negative word embedding values in SLDA as shown in Eq. (3).

$$\overrightarrow{we_{tfidf_d}} = \sum_{i \in W_d} (wt_{d,i} \times \overrightarrow{e_{d,i}}), \quad (3)$$

3.3 SLDA-based preference analysis

SLDA utilizes the word embedding vectors of news as the input of LDA to derive the semantic latent topics of each news article d, $\overrightarrow{nslt_d}$. Target user u's semantic latent topic $\overrightarrow{uslt_u}$ can be derived by averaging the semantic latent topics of news articles browsed by user u. The cosine similarities between the semantic latent topics of users and news can be calculated to derive target user u's SLDA-based interest score on news d, as defined in Eq. (4)

$$SLDARS_{u,d} = sim(\overrightarrow{nslt_d}, \overrightarrow{uslt_u}) \quad (4)$$

3.4 CSTM-based preference analysis

Collaborative Topic Modeling (CTM) combines LDA and MF to derive latent factors of users and documents [9]. The proposed Collaborative Semantic Topic Modeling (CSTM) integrates the semantic latent topic vectors of SLDA with the MF. A matrix R is built by U users' ratings on N items. Let $r_{u,d}$ denote the rating of user u on news d. $r_{u,d}$ is derived based on the browsing frequency of news d, as listed in Eq. (5).

$$r_{u,d} = \begin{cases} \log(1 + GBf_d \times Bf_{u,d}), & \text{if user } u \text{ had read the news} \\ 0, & \text{if user } u \text{ had not read the news} \end{cases}, \quad (5)$$

where GBf_d is the global browsing frequency of the news d for all users; and $Bf_{u,d}$ is the browsing frequency of the news d for user u. The rating scores are normalized into 0 to 5. Matrix R can be decomposed into two low rank matrices, X and Y, which represent the K-dimensional latent factors for users and news, respectively, by solving the optimization problem defined as Eq. (6):

$$\min_{x^*, y^*} \sum_{u,d} c_{u,d} (r_{u,d} - \hat{r}_{u,d})^2 + \lambda_x \|x_u\|^2 + \lambda_y \|y_d\|^2, \quad (6)$$

where x_u /$y_d$represents user u's/news d's latent factor; $\hat{r}_{u,d} = x_u^T y_d$ represents user u's predicted preference score on news d; λ_x and λ_y are regularization parameters; $c_{u,d}$ is a confidence parameter. $c_{u,d} = 1$, if $r_{u,d} > 0$; $c_{u,d} = \beta$, if $r_{u,d} = 0$.

Different from using random values in MF or employing LDA topic distribution in CTM as the initial values of the latent factors for items, we apply semantic latent topics $\overrightarrow{nslt_d}$ derived from the SLDA analysis to build the initial news latent factor of y_d. The generative process in CSTM is as follows:

(a) For each user u: set user latent factor $x_u \sim \mathcal{N}(0, \lambda_x^{-1} I_K)$.
(b) For each news d: set news latent factor $y_d = (\overrightarrow{nslt_d} + \theta_d)$ and $\theta_d \sim \mathcal{N}(0, \lambda_y^{-1} I_K)$.
(c) For each user-news pair, draw $r_{u,d} \sim \mathcal{N}(x_u^T y_d, c_{u,d}^{-1})$.

Let K be the number of latent topics and represent the dimension of latent factor. In the training process, we minimize the loss function by using the update equation defined in Eq. (7):

$$\begin{aligned} x_u &\leftarrow (YC_uY^T + \lambda_x I_K)^{-1}(YC_uR_u), \\ y_d &\leftarrow (XC_dY^T + \lambda_y I_K)^{-1}(XC_dR_d + \lambda_y \overrightarrow{nslt_d}), \end{aligned} \quad (7)$$

where C_u and C_d denote diagonal matrices; $R_u = (r_{u,d})_{d=1}^N$ for user u; $R_d = (r_{u,d})_{u=1}^U$ for news d; I_k is an $K \times K$ identity; Given $X = \mathbb{R}^{K \times U}$ and $Y = \mathbb{R}^{K \times N}$, we extract user latent factor x_u and news latent factor y_d by CSTM. The predicted CSTM-based user interest score is listed in Eq. (8):

$$CSTMRS_{u,d} = \hat{r}_{u,d} = x_u^T y_d \quad (8)$$

3.5 Ensemble model for user preference analysis

The ensemble model (EM) combines SLDA and CSTM. The ensemble interest score $EIS_{u,d}$ of user u on news d is expressed in Eq. (9).

$$EIS_{u,d} = (1-\alpha) \times SLDARS_{u,d} + \alpha \times CSTMRS_{u,d} \quad (9)$$

where the parameter α is used to adjust the relative importance between the SLDA-based and CSTM-based interest scores.

3.6 Recommendation adjustment

The proposed ensemble model is further integrated with the recommendation adjustment mechanism to derive the recommendation list. We adopt the most frequently pushed (MFP) and not frequently clicked (NFC) strategies [13] to adjust the recommendation list. The MFP strategy gives a lower push score to the news which has a higher push frequency but is not clicked by the target user. Let $PushFreq_{u,d}$ denote the recommendation frequency of news d for user u. The NFC strategy gives a lower push score to the news that is seldom clicked by global users or neighbors. The NFC push score of user u on news d, $NFCPushS_{u,d}$, is derived based on the click frequency generated by the global users and user u's neighbors. After calculating the MFP and NFC push scores for each news d, the harmonic mean is used to combine the two scores into the push score $PushS_{u,d}$ of user u on target news d.

We derive the final recommendation score $RS_{u,d}$ for user u on news d by using the harmonic mean of interest score $EIS_{u,d}$ and push score $PushS_{u,d}$, as defined in Eq. (10).

$$RS_{u,d} = \frac{2 \times EIS_{u,d} \times PushS_{u,d}}{EIS_{u,d} + PushS_{u,d}} \quad (10)$$

4. Experiment and evaluation

4.1 Dataset and experiment setup

We find the best predicting model in the offline experiments, and then conduct online evaluation on the target website: NIUSNEWS (https://www.niusnews.com/). In the offline experiments, the NIUSNEWS website provides 7394 news files and 2453 valid users. 75% of the data are used as the training set and the rest are for testing. We use the precision, recall and F1 metrics [18] to evaluate the recommendation performance and select the best parameter values of the proposed approach in the offline experiment phase. We evaluate the recommendation quality by utilizing the click-through rate (CTR) in online recommendation. CTR is the click frequency divided by the number of recommendations.

In offline experiments, we compare two baseline approaches including MF and CTM with two proposed approaches, including CSTM and EM. We also conduct online evaluation to compare our proposed approaches with the baseline CTM method.

- Recommendation based on Matrix Factorization (MF)
- Recommendation based on Collaborative Topic Model (CTM): The approach uses LDA latent topics of news as the initial news latent factors for Matrix Factorization.
- Recommendation based on Collaborative Semantic Topic Model (CSTM): The approach uses semantic latent topics of news as the initial news latent factors for MF.
- Recommendation based on Ensemble Model (EM): The approach combines SLDA and CSTM.
- Recommendation based on Ensemble Model with Recommendation Adjustment (EM-RA): The approach adjusts recommendation lists for online recommendation.

4.2 Model evaluation and parameter setting

Evaluation of parameters for SLDA: We utilize TFIDF as the input of LDA, while WE (word embedding with term weight) and WE_TFIDF as the inputs of SLDA to determine the parameters. We compare the performance under different latent topic number K. The result shows that the highest performance of WE_TFIDF-SLDA occurs at K = 140. WE_TFIDF-SLDA performs better than WE-SLDA, while WE-SLDA performs better than the baseline TFIDF-LDA. Hence, we use WE_TFIDF-SLDA with K =140 to denote the SLDA method.

Evaluation of parameters for CSTM: We compare different topic number (latent factor number) K on LDA for CTM and SLDA for CSTM. The result shows that MF and CTM achieve highest recommendation performance at K = 80; and the highest performance of WE-CSTM and WE_TFIDF-CSTM occur at K = 70.

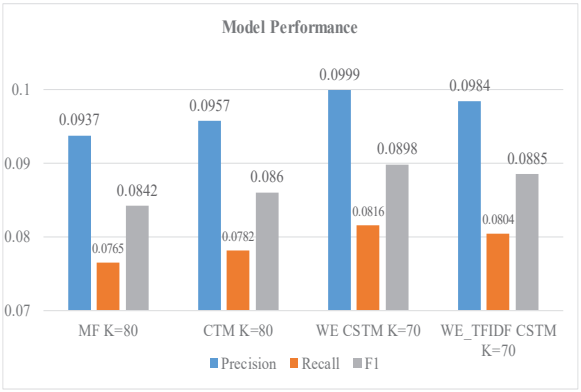


FIGURE 2. The comparison of CSTM models

Figure 2 shows that our proposed CSTM models perform better than the baseline MF and CTM models. This implies that considering semantic meaning (word embedding) in CSTM model can enhance the predictions of user preferences. The WE-CSTM method performs slightly better than WE_TFIDF-CSTM. Thus, we adopt WE-CSTM (K = 70) to denote the CSTM method in the following experiments.

Evaluation of parameters for EM: The proposed EM model derives user preferences by combining SLDA and CSTM. We compare different parameter values of α for combining the two methods SLDA and CSTM. We can obtain the highest recommendation performance when α (the weight of CSTM) is equal to 0.8. Hence, we employ this combination for the EM method in the following experiments.

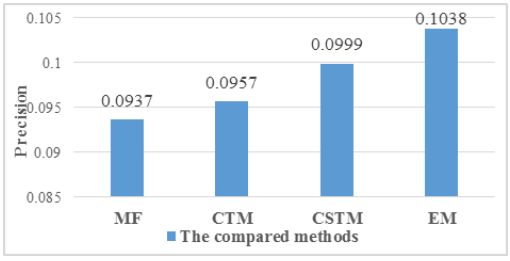
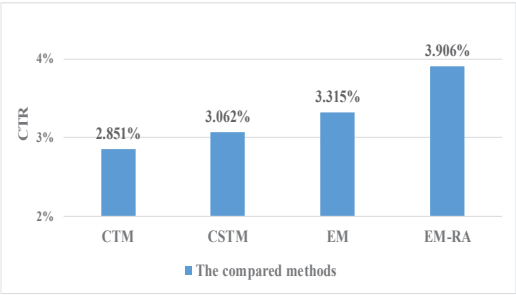


FIGURE 3. The comparison for the offline experiment

We compare different methods in offline experiments. Figure 3 shows that the EM approach performs better than CSTM approach, while CSTM performs better than the two baselines CTM and MF. The result demonstrates that the proposed CSTM model considering semantic analysis (word embedding) can enhance the recommendation quality. Moreover, the proposed ensemble model combining SLDA and CSTM models can further improve recommendation performance.

4.3 Offline and online evaluations of various methods



FGURE 4. The comparison for the online recommendation

In the online recommendation, we implement our proposed approaches on the NIUSNEWS (https://www.niusnews.com) website to conduct online news recommendation. We compare the baseline CTM with our proposed methods, including CSTM, EM and EM-RA. The online evaluation result is shown in FGURE 4. Our proposed CSTM model has higher recommendation quality (CTR) than the baseline CTM. This implies that the proposed CSTM considering semantic analysis with word embedding can more effectively extract the latent features of news articles. Moreover, the ensemble model can improve online recommendation quality by integrating different aspects of user preferences.

Moreover, our proposed EM-RA method considering recommendation adjustment (RA) can perform better than the methods without considering RA mechanism such as CTM, CSTM and EM. This indicates that dynamically adjusting the recommended news articles helps to increase the chance of recommending interested news to users.

5. Conclusions

In this work, a novel news recommendation approach is proposed. We employ semantic vectors of word embedding to design a Semantic LDA (SLDA) model. Then we build a semantic latent factor for Matrix Factorization to design a Collaborative Semantic Topic Model based on combining CTM with SLDA. We also integrate SLDA with CSTM models to design an ensemble model (EM) for analyzing users' preferences. The evaluation shows the proposed methods can improve the recommendation quality and obtain better performance than typical CTM and MF approaches. The proposed method EM-RA considering recommendation adjustment (RA) can perform better than the methods without considering RA mechanism. Future work will investigate online interest analysis to adjust the online interest score by considering the news articles currently browsing.

Acknowledgment

This research was supported by the Ministry of Science and Technology of Taiwan under Grant No. 105-2410-H-009-033-MY3.

References

[1] L. Li, L. Zheng, F. Yang, and T. Li, "Modeling and broadening temporal user interest in personalized news recommendation," Expert Systems with Applications, Vol. 41, No. 7, pp. 3168-3177, 2014.

[2] Y. Wang and W. Shang, "Personalized news recommendation based on consumers' click behavior," Proceedings of the 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2015, pp. 634-638.
[3] M. Zihayat, A. Ayanso, X. Zhao, H. Davoudi, and A. An, "A utility-based news recommendation system," Decision Support Systems, Vol. 117, No. pp. 14-27, 2019.
[4] M. Yajun and L. Sheng, "Application of LDA-LR in Personalized News Recommendation System," Proceedings of the 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), 2018, pp. 279-282.
[5] M. Kompan and M. Bieliková, "Content-based news recommendation," Proceedings of the International Conference on Electronic Commerce and Web Technologies, 2010, pp. 61-72.
[6] M. K. Najafabadi, M. N. r. Mahrin, S. Chuprat, and H. M. Sarkan, "Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data," Computers in Human Behavior, Vol. 67, No. pp. 113-128, 2017.
[7] Y. Koren and R. Bell, "Advances in collaborative filtering," in Recommender systems handbook, ed: Springer, 2015, pp. 77-118.
[8] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," IEEE Computer, Vol. 42, No. 8, pp. 30-37, 2009.
[9] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011, pp. 448-456.
[10] Y. Li, M. Yang, and Z. M. Zhang, "Scientific articles recommendation," Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, 2013, pp. 1147-1156.
[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, Vol. 3, No. Jan, pp. 993-1022, 2003.
[12] R. Das, M. Zaheer, and C. Dyer, "Gaussian lda for topic models with word embeddings," Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015, pp. 795-804.
[13] D. R. Liu, K.-Y. Chen, Y.-C. Chou, and J.-H. Lee, "An Online Activity Recommendation Approach Based on the Dynamic Adjustment of Recommendation Lists," Proceedings of the Advanced Applied Informatics (IIAI-AAI), 2017 6th IIAI International Congress on, 2017, pp. 407-412.
[14] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in The adaptive web, ed: Springer, 2007, pp. 325-341.
[15] P. Lops, D. Jannach, C. Musto, T. Bogers, and M. Koolen, "Trends in content-based recommendation," User Modeling and User-Adapted Interaction, Vol. 29, No. 2, pp. 239-249, 2019.
[16] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proceedings of the National academy of Sciences, Vol. 101, No. suppl 1, pp. 5228-5235, 2004.
[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," Advances in Neural Information Processing Systems, 2013, pp. 3111-3119.
[18] R. Baeza-Yates and B. Ribeiro-Neto, Modern information retrieval: ACM press New York, 1999.