Genre Prediction to Inform the Recommendation Process

Nevena Dragovic
Computer Science Department
Boise State University
Boise, ID, USA
nevenadragovic@u.boisestate.edu

Maria Soledad Pera
Computer Science Department
Boise State University
Boise, ID, USA
solepera@boisestate.edu

ABSTRACT

In this paper we present a time-based genre prediction strategy that can inform the book recommendation process. To explicitly consider time in predicting genres of interest, we rely on a popular time series forecasting model as well as reading patterns of each individual reader or group of readers (in case of libraries or publishing companies). Based on a conducted initial assessment using the Amazon dataset, we demonstrate our strategy outperforms its baseline counterpart.

CCS Concepts

•**Mathematics of computing** → **Time series analysis;**
•**Information systems** → *Recommender systems;*

Keywords

Prediction; Genre; Books; Time Sequence; ARIMA

1. INTRODUCTION

Books, which constitute a billion dollar industry[1], are the most popular reading material among all generations of readers, both for leisure and educational purposes. Hundreds of thousands of books of different types (e.g., paperback and e-books) and styles (e.g., fiction and non-fiction) are published on a yearly basis, giving readers a variety of options to choose from. Book recommendation systems, which are meant to enhance the decision making process, can help users by identifying, among the sometimes overwhelming number of diverse books, the ones that best suit their interests and preferences. These recommenders are not exclusively designed to aid *individuals* in their quest for reading materials. They can also improve the decision making process for *libraries*, by suggesting what books to buy in order to maximize the use of library resources by their patrons, and *publishing companies*, by advising which books to publish in order to maximize revenue. To better serve stakeholders, recommenders must be able to predict interest and needs. However, given that preferences may alter over time for different readers, the *time* component is important and crucial to consider in the prediction process [5].

There are many avenues that can be explored from a time-sensitive stand point in order to generate better recommendations, including user-generated ratings, reviews or books metadata. One of them, which is often overlooked and is the focus of our paper, is genre. By its definition, genre (e.g., drama, comedy) is a category of literary composition, determined by literary technique, tone, content, or even length. While genre has been studied as a part of the recommendation process [4], the influence of its distribution over time on suggesting suitable books for individual or group of users (as in the case of libraries and publishing companies) has not been explored. Change of genre in time is a significant dimension to improve the genre prediction process. This can consequentially influences the process performed by book recommenders since it provides the likelihood of reader(s) interest in each genre based on its occurrences at a specific point of time, not only the most recent or the most frequently read one. As an answer to this need, we propose a genre prediction strategy that examines genre distribution over time and applies time series analysis models. The goal of our strategy is to discover different areas of users' interests, not only the most dominant ones.

From the users' point of view, explicitly including time based analysis to inform the recommendation process will lead to relevant suggestions that satisfy their specific reading needs. Finally, from the commercial point of view, the benefit would be in understanding the influence of reading patterns on decisions about what genre should be published or acquired in a given point of time.

2. RELATED WORK

A considerable number of studies examine the importance of book genre on readers' activity [1, 2]. However, to the best of our knowledge, research based on past genre distribution coupled with time series analysis to influence the recommendation process has not been conducted. As presented in [4], genre is used as a data point to inform a cross-domain collaborative filtering approach that recommends books based on users' genre preferences. You et al. [6] propose a clustering method based on users' ratings and genre interests extracted from social networks to solve the cold-start problem affecting collaborative filtering approaches. Unlike the proposed methods, we use time series to predict the genres most likely currently of interest to each individual user, which can further enhance the book recommendation process.

3. METHOD

Predicting genre to inform the recommendation process, regardless of the major stakeholder (a reader, library or publishing company), involves examining genres considered in the past, either read by specific users or purchased by

[1] http://goo.gl/GMn8Nc

customers. While a simple genre distribution analysis yields probabilities or weights that determine the most favored genres, it lacks the ability to consider genre preference evolution over time. To overcome this drawback, we propose a time-based genre examination which requires information on reading activities among readers. As a first step to our proposed strategy, we explore reading activity of a user to obtain the distribution of his/her genre interest during continuous periods of time. IN every period of time, for each genre we calculate a significance score that captures its importance by considering a number of books read of that genre in that period of time. Thereafter, to explicitly consider the change of genre preference distribution over time, our genre prediction strategy takes advantage of Auto-Regressive Integrated Moving Average[2] (ARIMA). We selected ARIMA since it is one of the most popular models that uses time series for prediction purposes.

By using ARIMA we are able to determine a model tailored to each genre distribution to predict its importance for the corresponding user in real time based on its previous occurrences. Note that each predicted genre importance score is based on: its occurrences in the past, a specific time when it occurred and its importance for a specific user. To define length of time periods used by ARIMA, we used information from a recent study done by Pew[3] on reading habits in the USA, we establish one month long "windows" of time in which each user is expected to read at least one book, so our strategy uses 1 month time frames from the the first book log (either bookmarked or rated book) to last.

4. INITIAL EVALUATION

Framework. To validate the performance of our proposed time-based genre prediction strategy, we selected a subset of the Amazon/LibraryThing[4] book dataset. Since the **dataset** does not always include genre as a part of the provided metadata, we extended it by including genre information from the Library of Congress[5]. We used 1214 users[6] along with the books they rated or reviewed. To **quantify** the assessment, we applied Mean Average Error (MAE), Accuracy and Kullback-Leibler (KL) divergence [3]. MAE estimates the difference between the predicted genre importance and the ground truth, i.e., genre distribution for a user at a given time, whereas Accuracy applies a binary strategy that reflects if the predicted genres correspond to the ones read by a user in a given period of time. KL divergence measures how well a distribution q generated by a prediction strategy approximates to distribution p, the ground truth. In establishing the **ground truth** for each user considered for evaluation purposes, we adopted the well-known *N-1* strategy, such that the genre of the books rated by a given user U in the N time frame are treated as "relevant" genres for U, and the genre of the books rated in the previous *N-1* windows are used for training U's genre prediction model. As a **baseline** of our initial assessment, we use a traditional prediction strategy that considers the proportion of occurrences of each genre based on data collected over *N-1* periods of time to estimate the importance of each genre for a given user on the current, i.e., N, time period.

Results. As shown in Table 1, for N=11[7] outperform the baseline. KL divergence scores showcase that genre distribution predicted using time-series approach better approximates to the ground truth. Furthermore, the probability of occurrence of each considered genre is closer to the real values when the time component is included in the prediction process. As a further assessment, we observed the differences in genre predictions among users who read different number of distinct genres. For users who read only one to two genres, the time-based prediction strategy does not perform better than the baseline. However, if a user reads three or more genres, our time-based genre prediction strategy outperforms the baseline in all three metrics. This is not surprising, given that it is not hard to determine area(s) of interest for a user who constantly reads only one or two book genres, which is why the baseline performs as good as time-based prediction strategy. Given that users that read 3 or more genres represent 91% of the users in our sampled dataset, the proposed strategy provides significant improvements in predicting preferred genre for the vast majority of readers.

Table 1: Evaluation using the Amazon dataset

	MAE	KL	ACC
With Time Series	0.143	0.623	0.870
Without Time Series	0.144	0.663	0.826
With Time Series (3+ genre)	0.138	0.660	0.857
Without Time Series (3+ genre)	0.146	0.720	0.810

5. CONCLUSIONS

In this paper, we described our efforts in developing a time-based genre prediction strategy that can better inform the recommendation process. The novelty of our approach consists of incorporating an explicit time component to generate genre distribution. To the best of our knowledge, this is the first time that the well-known time series ARIMA model is used to predict book genre of readers' interests. The described strategy provides successful predictions and outperforms the baseline for 77% of users based on the presented initial evaluation, while for the remaining users it provides predictions comparable to the baseline. Because of the scope of this paper, the conducted evaluation showcases the genre prediction performance for a single user, while we still need to conduct further assessments in terms of quantitatively determining the degree to which the proposed strategy (i)provides successful genre predictions for libraries and (ii)publishing companies and influences the recommendation process to assist all three stakeholders.

6. REFERENCES

[1] P. Afflerbach. The influence of prior knowledge and text genre on readers' prediction strategies. *Journal of Literacy Research*, 22(2):131–148, 1990.

[2] J. Anderson, A. Anderson, J. Lynch, and J. Shapiro. Examining the effects of gender and genre on interactions in shared book reading. *Literacy Research and Instruction*, 43(4):1–20, 2004.

[3] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

[4] S. Sarawagi and S. H. Nagaralu. Data mining models as services on the internet. *ACM SIGKDD Explorations Newsletter*, 2(1):24–28, 2000.

[5] J. Wang and Y. Zhang. Opportunity model for e-commerce recommendation: right product; right time. In *ACM SIGIR*, pages 303–312, 2013.

[6] T. You, A. N. Rosli, I. Ha, and G.-S. Jo. Clustering method based on genre interest for cold-start problem in movie recommendation. *JIIS*, 19(1):57–77, 2013.

[2] http://goo.gl/Dhzcg7

[3] http://goo.gl/BAUQK4

[4] http://goo.gl/drH0yF

[5] https://www.loc.gov/

[6] In our initial assessment, we considered Amazon users who provided ratings for at least 35 books.

[7] We empirically verified that for 6<N<11 the results are comparable to the ones for N=11