Knowledge-Based Systems 105 (2016) 134–146

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Group-based Latent Dirichlet Allocation (Group-LDA): Effective audience detection for books in online social media

Peng Zhang[a,b], Hansu Gu[c,**], Mike Gartrell[d], Tun Lu[a,b,*], Dayi Yang[a,b], Xianghua Ding[a,b], Ning Gu[a,b]

[a] School of Computer Science, Fudan University, Shanghai, China
[b] Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China
[c] Seagate Technology, Longmont, CO, USA
[d] University of Colorado Boulder, Boulder, CO, USA

ARTICLE INFO

Article history:
Received 18 January 2016
Revised 28 March 2016
Accepted 7 May 2016
Available online 10 May 2016

Keywords:
Social media
Book recommendation and marketing
Audience detection
Group-LDA

ABSTRACT

Most current book recommendation and marketing strategies in online social media are implemented by creating topics or posting advertisements for the brand. They do not precisely target the audiences who are interested in these books, so the recommendation or marketing quality is not guaranteed. In order to solve this problem, we propose an effective audience detection method based on Group-based Latent Dirichlet Allocation (Group-LDA) in order to precisely detect book audiences. Group-LDA is a new probabilistic topic model derived from Latent Dirichlet Allocation (LDA), which introduces a new latent concept of *group* to describe the topic relevance among documents by incorporating book module and book chapter information into the model. Group-LDA is evaluated on *Weibo.com* with fifty popular books randomly sampled from the reading channel on *Douban.com*. According to the evaluation results, Group-LDA can effectively detect different types of readers for most categories of books. It outperforms LSA, LDA, author-topic model (ATM) and some other collaborative filtering methods in terms of precision, recall, F1-score and MAP for book audience detection.

1. Introduction

Book recommendation and marketing in online social media such as Weibo, Twitter, and Facebook has become increasingly important. According to the *Global Trends in Publishing 2014*, the size of the market for books is larger than the market for other content-based creative products such as movies, televisions, music, and games [1]. Thanks to the real-time, user-generated, and interactive features of online social media [2], it is considered effective to brand books in online social media. However, most current recommendation and marketing strategies for books involve creating topics or posting advertisements for the brand, which do not target the precise audience interested in these books and cannot guarantee desired recommendation and marketing quality. These strategies are considered passive because they ignore the fact that users who have already shared their opinions and interests can be accurately targeted. Therefore, we seek to address the problem of audience detection for books in online social media.

Audience detection for books in online social media is a new and important problem. On one hand, due to the large user base and the purposes underlying the use of online social media platforms, explicit user reading history is often missing, and instead personal interests and opinions on certain topics are posted. The lack of reading history falls outside of the definition of problems solved by many collaborative filtering based approaches. Therefore, audience detection should be defined based on the content of user messages (we utilize microblog posts as motivation examples in this paper), because these messages are the direct and real-time representation of users' preferences, sentiments, and perceptions, which have already been utilized as a valuable data source for product and article recommendation [3,4]. On the other hand, the diversity of user topics in online social media [5] adds a significant variety in content which weakens the explicit correlation between the content of the book and user messages. This characteristic makes the problem different from traditional information retrieval problems which rely on the close relevance of query and documents. Audience detection, which aims to identify readers of books using content-based analysis and the incorporation of diverse user generated information, has become a new and

* Corresponding author at: School of Computer Science, Fudan University, Shanghai, China.
** Corresponding author.
E-mail addresses: zhpll@126.com (P. Zhang), guhansu@gmail.com (H. Gu), lutun@fudan.edu.cn (T. Lu).

http://dx.doi.org/10.1016/j.knosys.2016.05.006
0950-7051/

increasingly important problem. It has direct application in book marketing and dramatically enhances its accuracy.

Audience detection for books is a challenging task. Traditional topic modeling approaches including Latent Dirichlet Allocation (LDA) [6] and author-topic model (ATM) [7] cannot be used directly to solve the problem. According to Adler and Van Doren [8], there are four levels of reading including elementary reading, inspectional reading, analytical reading and syn-topical reading. In the former three levels, the readers generally focus on a book's whole content, while in the last level, the readers only emphasize on one or more chapters which are related to a special topic. Thus in this paper, we classify a book's readers into three categories including the users who are interested in the book's whole content (general readers), the users who are interested in one of the book's chapters (special readers) and the users who are interested in many of the book's chapters (professional readers). As a book's summary describes the key points of the book's whole content, general readers can be mined by evaluating the topic similarity between users' microblog posts and the summary. Similarly, the special readers are also not difficult to detect through measuring the topic similarity between users' microblog posts and each chapter summary of the book. While for the professional readers, they focus on the combination of some chapters' content, which results in that the topics of their microblog posts are not similar to ones of the book summary nor any chapter summary. Thus these potential readers who are proved accounting for a very large proportion among a book's potential readers in our experimental evaluation cannot be detected by the original LDA model. The reason for the problem above is that there is always some topic relevance between a book's chapters, and users prefer to talk about many related chapters together through their microblog posts. For example, in a book on algorithm, some operations in the chapter "string operation" are based on the algorithms in chapters "sort" and "search", which results in that the former is relevant to the latter chapters, and these chapters co-occur frequently in users' microblog posts. Another example is the three chapters "struts", "spring" and "hibernate" which constitute together a module SSH in a book focusing on Java programming. Unfortunately, the original LDA model extracts topics from documents while ignores the latent relevance. The group of co-authors existing in ATM provides an idea to represent the topic relevance between the book's chapters, while each author is drawn uniformly from co-authors, we argue in order to model a book, different chapters in a module should carry different weights as shown later in our work.

To address the challenges above, we propose a new probabilistic topic model named Group-based Latent Dirichlet Allocation (Group-LDA) by introducing the concept of *group* into the LDA model in order to represent the topic relevance between documents. Based on our proposed model, professional readers can be detected by evaluating the group similarity between their microblog posts and each chapter summary of the book. This paper makes the following major contributions:

- We identify and define the problem of audience detection for books in online social media. To the best of our knowledge, this is the first work on audience detection for books in social media according to the unique characteristics of books.
- We propose a modified probabilistic topic model incorporating the variety of user generated content.
- We evaluate the performance of our proposed audience detection method on a dataset collected from both *Weibo.com* and *Douban.com*.

The rest of the paper is organized as follows. In Section 2, we present an overview of related work. In Section 3, we introduce preliminary analysis results showing the quality of book recommendation and marketing through open topics and how a user's reading preferences are influenced by other users in social media. In Section 4, we define the problem of book recommendation and describe the architecture of our solution. The Group-LDA model is described in Section 5. Evaluation results are presented in Section 6. Finally, we conclude the paper and describe future work in Section 7.

2. Related work

With the explosive growth of the Internet, the problem of information overload has become more and more critical. For enterprises and organizations, finding consumers of interest from a large population of users is quite a challenge as a result of the rapid growth of the user base and their diverse activities on the Web [9,10]. Individual users are confronted with the challenge of abundant information that is difficult to filter and transform into useful insights.

Information filtering (IF), which aims to help users to obtain relevant content from large amount of information, is an effective method to alleviate the problem of information overload. In the research area of IF, the most representative systems are web data extraction systems [9,11] and recommender systems [12–14]. Specifically, online recommendation and marketing predicts user interest and consumer demand by analyzing and comprehending the online behavior of users, which provides an effective solution for information filtering. However, much recent research on recommendation methods and marketing strategies is mainly based on the collaborative filtering (CF) or its improved algorithms, such as the normal recovery collaborative filtering approach for personalized web service recommendation [15], the incremental CF recommender based on the Regularized Matrix Factorization [16] and the typicality-based collaborative filtering recommendation method named TyCo [17], in which a user's interest in a particular product is evaluated using her/his historical profiles. In order to improve the recommendation quality of documents such as scientific articles, the collaborative topic regression (CTR) model [18] which combines traditional collaborative filtering with topic modeling has been proposed. Previous studies have shown that CTR works well compared to traditional matrix factorization methods [19,20]. However, the CF has defections of cold start [21] and data sparsity [22] which reduce its practical performance.

The emergence of online social media supplies us with abundant information for extracting user interest and preference, as well as an increasingly efficient platform for implementing recommendation methods and marketing strategies. Sociology research indicates that a user's interest and consumer behavior can be influenced by her/his neighbors [23,24], especially by individuals who have similar interests or a high level of expertise. Furthermore, there are often communities in social media composed of users with the same interest. A user's consumer behavior can be greatly affected by other users in the same community. Social media therefore provides a direct, real-time, and multi-user dissemination platform for product recommendation and marketing. Existing social media recommendation methods mainly include individual-centered methods and product-centered methods. The former, which has attracted much recent research effort, focuses on individual users and the design of a personalized recommendation list for each user according to her/his personal information, social context, and interaction patterns in social media [25–27]. While the product-centered recommendation method seeks to design a recommendation system that uses audience detection, advertising, and social network promotion for each product according to its particular features, such as the studies on extracting users' attitudes to products [28], forecasting future outcomes [29] and customer identification [30]. Unfortunately, research on such methods nowadays is scant, especially for books. Although a few