Library Book Recommendations Based on Latent Topic Aggregation

Shun-hong Sie

Ph.D student of Graduate Institute of Library & Information Studies, NTNU, modify@ma37.hinet.net

Jian-hua Yeh

Department of Computer Science and Information Engineering, Aletheia University, jhyeh@mail.au.edu.tw

Abstract. During recent years, how to provide personalized services has become an important research issue in library services. The libraries provide more and more personalized services such as customized web interface and reading suggestions. In the traditional approaches, the features of the books that a reader likes are used to construct the profile of the reader to support recommendation of books such as query keywords. But with the fact of the huge holdings in the libraries, the librarians need to effectively help the readers to find the books of interest. Collaborative filtering (CF) is a way to make it possible by use patron’s circulation logs which contain their borrow history as favorite readings. In this paper, we first use Latent Dirichlet Allocation to find the latent topics existing in the circulation logs, then we combine patron reading histories with the generated latent topics to produce a suggestion list for the patron. With the elaborated experiments demonstrated in this paper, it showed good results from the volunteers’ feedback.

Keywords: book suggestion, latent topic, library service, collaborative filtering

1 Introduction

The librarians provide more and more personalize services in the recent years, for example, the personalized web and some reading suggestions. In traditional personal book suggestions, it is done by adopting the books’ features compared with personal profiles or query keywords, trying to form suggestion lists for library readers. The librarians should provide more information to the patrons to fit their preferences by recommending suggestion book lists, and help the patrons to discover books which they may never seen but hold in the library.

According to the analysis of a patron’s reading history or some query keywords given by the patron, the library is able to provide some new information or books
about it. But sometimes the patron’s query keywords or reading history might not be effective enough and may cause lot of time to identify the patron's need. Since the library contains a huge amount of holdings which can only be accessed or searched via on-line catalog or browsed stock by stock. So most people do not know the holding situation which may contain some information they need, and some of interesting books outside the hot topics will not be seen in library on-line catalog. On the other hand, basing on patron’s search skill or behavior might cause some important resources or interest for patrons not found or unseen.

Base on use statistics on library circulation counting indicate that most books are utilized by very few patrons [1]. And these collaborative approaches might tend to recommend popular titles which might have high rating on it, perpetuating homogeneity in reading choices. It might impossible for a collaborative approach to recommend items that have no one rated on it or just purchased into the library.

Using data mining technologies on analyzing huge patrons’ reading histories should be helpful to the librarians to recommend holdings for the patrons. In this research, we focus on building a model to render reading list suggestions for patrons through collaborative filtering. The Latent Dirichlet Allocation (LDA) model is adopted to find the latent topics in the library circulation logs, then the render suggestion list is produced by matching patron barrow history.

This paper aims at developing an personalized reading profile construction process to be used on personalized suggestion. Before developing the construction process, all the library’s circulation logs should be fetched and processed. We use the National Taiwan Normal University library’s circulation logs for our experiment, all the information with personal security concerns or private data was removed. We try to find a user’s suggestion reading list which he/she may be interested in. Finally we use K-fold experiments with human questionnaire to evaluate our suggestion list. The experiment results shows that our system had good performance on users satisfaction.

2 Relate Works

In this section, we focus on the discussions about the relationships on catalog number and latent topic discovery which will be used to find the latent topics, and these topics are used to compare user profiles in order to make suggestion readings lists for them. The catalog number is used to represent knowledge organization in the library, not only for librarians but also for patron to search holdings in the library.

2.1 Catalog Number

Catalog number is a way of organizing library holdings. In most libraries, holdings are arranged according to subject-oriented classification schemes. The decimal system scheme which used in the National Taiwan Normal University library is named “Chinese Library Classification”. It is based on Dewey Decimal System and is modified to fit the Chinese holding environment. It is a hierarchy classification.
system to represent knowledge from top level to bottom level, and group different holdings together which may contain different title or writer by different author.

2.2 Latent topic discovery

For the researches in latent topic discovery, most of the research focuses aim at topic detection in text data by using term distribution calculation among the documents. Several important algorithms were developed, including Latent Semantic Analysis (LSA)[3], Probabilistic Latent Semantic Analysis (pLSA)[2], and Latent Dirichlet Allocation (LDA)[4]. LSA is one of the semantic analysis algorithms which differs from traditional term frequency-inverse document frequency (TF-IDF) model. pLSA model is proposed to overcome the disadvantage found in by LSA model, trying to decrease the degree of computation by using probabilistic approach. pLSA and LSA try to represent the original document space with a lower dimension space called latent topic space. The algorithm of Latent Dirichlet Allocation (LDA) is more advantageous since LDA performs even better than previous research results in latent topic detection. In fact, LDA is a general form of pLSA, the difference between LDA and pLSA model is that LDA regards the document probabilities as a term mixture model of latent topics. Girolamin and Kaban [5] shows that pLSA model is just a special case of LDA when Dirichlet distributions are of the same.

The goal of our recommendation systems is to give personalized recommendation on items to users. Typically the recommendation is based on the former and current activity of the users, and the metadata about users and items if available.

3 Proposed Method

Call numbers are used to describe a book, it consists of a number, and is also used to indicates the location on the shelf where the item can be found. A call number is composed by catalog number, author number and volume number, sometimes it might add a special mark for labeling that it’s a special collection. For example, the reference book will be labeled by adding an “R”. This represents an item's subject matter and the different collection which have similar call number, that is, they should have similar subject. In this work, we use patron’s circulation records for the suggestion of system data source.

For each patron, $u$, who’s circulation records could be represented by $u = [b_1, b_2, b_3, ..., b_n]$, and for each book, $b$, which could be represented by call number $c$. Thus patron’s circulation records could be formatted as $u = \{c_1, c_2, ..., c_n\}$. The patrons’ circulation records could be treated as a sparse matrix. But the call number represents a subject, if we just use call number on the records, the recommendation list may not be correctly generated due to a narrow subject. In this work, the researchers decide to use the top subject, which means just fetch the first 4 number, by different weight of sum adjustment to represent the subject similarity. We define the PreMatch function to describe this situation. The number $i,j$ represent two different call number, and PreMatch($i,j$) return the a weight value $w$, see Eq. 1.

$$score w = \frac{PreMatch(i,j)}{|i|}$$ (1)
The researcher uses LDA model to find the hidden subjects. Here we propose to use LDA to compute the circulation matrix and identify the latent topics which represent the subjects by call number. Finally, we use cosine similarity and adopt hierarchical agglomerative clustering (HAC)[6,7] algorithm to classify the border subject. The output of a typical HAC algorithm is a classification with hierarchical structure, and there exists the need to merge the nearby nodes to generate subject groups. The result can be treated as a list profile with subjects represented by the composition of a lot of call numbers. By using the cosine similarity to compare the patron’s historical circulation records and the list profile generated by the above steps, it is able to generate a possible subject list which a patron might be interested in.

4 Experiments and Evaluations

The following sections present empirical results obtained from evaluating our approach. We use the circulation log to test our method output, finally building an online evaluation web page to use human judge. The result shows our system has made good suggestion holding list.

4.1 Data Sets

The data range is from 2006 Aug. to 2007 Jun. The data set which contains 137400 records, all personal private information was removed. For our evaluation, we divide the entire data set into 12 subset by month.

Table 1. show the catalog number distribution which show the catalog number 800 is most user barrow than the others, next one is 300. Its show maybe hidden a static information between the user and suggestion books, maybe most person like to read it. But the main catalog number 800 is a very big scope for user to read, it may contain lot of sub catalogs in it, which mean it could contain all kind of literature just like novels and poems. We hope the suggestion result should be extract match full catalog number as we can, in order to suggest the book in a limited scope that user will interest in.

<table>
<thead>
<tr>
<th>Catalog number</th>
<th>000</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>0.042</td>
<td>0.064</td>
<td>0.030</td>
<td>0.230</td>
<td>0.052</td>
</tr>
<tr>
<td>Catalog number</td>
<td>500</td>
<td>500</td>
<td>600</td>
<td>700</td>
<td>800</td>
</tr>
<tr>
<td>%</td>
<td>0.033</td>
<td>0.033</td>
<td>0.065</td>
<td>0.084</td>
<td>0.256</td>
</tr>
</tbody>
</table>

4.2 Experiment setup and result

In order to evaluate our framework, we conducted experiment real data provided by National Normal University Library circulation log. These data contain patron id, book id and barrow date, but no catalog number in it. In order to perform our
method, we have to enrich data by find the catalog number, title, and author information from the WEBPAC, using crawler to fetch. All the error data which might be miss catalog number or patron id were removed. These error data may be cause by the patron was leave school or the book was miss so the bibliography data was removed.

We conducted 12-fold cross-validations for the experiment. This means that we randomly divided sets of all examples into 12 sets of examples: 11 of these were used for learning and one was used for testing. Then the data set for testing was rotated for 12 times in order. For each time, we randomly selected 100 users from logs to evaluate suggestion result compare with baseline trivial suggestion by using probability distribution. As a result, we can measure the performance for suggestion result by this experiment using recall and precision value. As can be seen from Table 4, our method have 0.7 recall, 0.72 precision on average, and stable result among the different data sets, beside the moth Aug, Jun and Feb which are the vocation, less patron barrow the books cause less information output.

In order to examine our method in the real situation, we develop an on line questionnaire system to collection user’s feedback. This system will output twenty suggestion records mix from our method and base line, each record provider book jacket, summary, author and catalog number information. Twenty-five persons’ feedback which total have 500 suggestion result were be collected as Table 2 show. Our system had got 60% user satisfy, it could render a good suggestion list to patron. In the other words, our suggestion method could satisfy user interest, user maybe barrow these book next time or read it.

<table>
<thead>
<tr>
<th>Table 2. User on-line evaluation result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Our method</td>
</tr>
<tr>
<td>Base line</td>
</tr>
</tbody>
</table>

5 Conclusion and future work

In this paper, we discuss use Latent Dirichlet Allocation (LDA) to extract latent topics and renderer the suggestion book lists. Although it perform a good result, but still need to enhance. According to the patron response, we have summary follow issue:

1. Suggestion result should include the other topic which user might be interest. Not only to suggestion the top N in output suggestion list, but also need to choice some of book which might be sort at last.
2. Sometime the provider information which contain book jacket or book summary, might be effect user’s interest. If we provider extra information about suggestion list, it might cause user change mind and barrow these book.
Because of limit by the system circulation log, each record does not have return date, so can not use barrow periods be a weight to measure patron’s reading time. We can not trace the change of user interest, if he reading a book for long time, might be he interest in than other which reading time is shortly. If we can gather more information or user feedback, the suggestion result might become more exact by follow user interest change, it could suggest the new scope that user will interest but had not discovery.

We plan to further validate our findings on the National Taiwan Normal University library WEBPAC system. We will pursue this method by integrating a “book suggestion” functionality into library website. To enhance the personalize function and provider more customization service, just like new book suggestion or journal suggestion. On the more theoretical side, future work will consider when data is grown up, an on line system which many patron active log on it would be a big challenge and need to solve.

6 Acknowledgments

The authors acknowledge with National Taiwan Normal University library for making their data available. And books.com(http://www.books.com.tw) for providing their books summary and book jackets information. Thanks for volunteer help us to evaluation our suggestion results.

7 References