2019

LIBRARY MANAGEMENT SYSTEM WITH TOPIC MODELLING AND ITS ADAPTABILITY TO OPEN AND DISTANCE LEARNING LIBRARIES

Babafemi RICHARD Adebayo
radebayo@noun.edu.ng

Follow this and additional works at: https://digitalcommons.unl.edu/libphilprac
Part of the Library and Information Science Commons

https://digitalcommons.unl.edu/libphilprac/2709
LIBRARY MANAGEMENT SYSTEM WITH TOPIC MODELLING AND ITS ADAPTABILITY TO OPEN AND DISTANCE LEARNING LIBRARIES

By Babafemi Richard Adebayo
National Open University of Nigeria

ABSTRACT

The adoption of Integrated Library Systems (ILS) became prevalent in the 1980s and 1990s as libraries began or continued to automate their processes. These systems enabled library staff work, in many cases, more efficiently than in the past.

However, these systems are restrictive and have thus undergone changes over the years, making processes more efficient. One area of improved capabilities is that of “search”, which in this paper, builds on integrating topic modeling as a new feature in modern integrated library systems in open and distance learning institutions. Users can now partake and explore new ways of resolving text classification and data exploration problems on a typical library management.

This aims also at improving book search, browse and help in book-selection decision making.

Keyword: topic modelling, integrated, library, LDA

INTRODUCTION

The use of libraries has grown tremendously in the last decade. Its processes, such as acquisition, cataloguing, shelving and the general management of information has evolved through the years, in aspects of digitisation and in fact knowledge management. All these would not have happened if not for the need for Librarians to make work easy for themselves and indeed the library users. This is one of the reasons Younis (2012) made aware that library users encounter problems when finding, borrowing, localising, renewing the borrowing, queuing for books. For most of the problems, solutions or ongoing researches are on to solve the problems. However, one area still short of indepth research, is the area of integrating topic modelling into library management systems. Topic modeling is a kind of a probabilistic generative model that has been used widely in the field of computer science with a specific focus on text mining and information retrieval in recent years (Liu et. al, 2016).

In recent years, focus have shifted by open and distance learning institutions to electronic books, hence the large tons of data in the information space. This ubiquitousness portends a challenge for librarians and most especially users in skimming through electronic data, in form of books, in finding their target information. It is safe to say, “you do not judge a book by its cover” or perhaps its table of content or even abstract. If a user cannot establish the
content of say, a 1000-paged book, for instance, by going through the table of content, abstract or book cover, then how?

The traditional means adopted by most users is to scan through each of the “suspected” books, before finally settling on one or two. This exercise can be tedious, tasking, time-consuming and in fact, sometimes, ineffective.

The digital age has exposed the minds of users into an unending world of possibilities, thereby, seeking new and more explorative ways of solving issues in library usage. One of these is topic modeling.

Topic modeling seeks to view a document as a probability distribution over a fixed vocabulary. As an example, Table 1 (The top five most frequent words from three topics) illustrates three “topics” that were discovered in a corpus, including “Library,” “Publishing,” and “Computation” (Blei 2012). Liu et. al (2016) explains as shown in Table 1, the probabilities of each word in a “topic” were sorted in the descending order. The top five most frequent words reflect the related concepts of each “topic”: “Topic 1” is about a library, “Topic 2” is about publishing, and “Topic 3” is about computation. In short, each “topic” is a mixture of “words” in a vocabulary. Similarly, in topic modeling, each document is a mixture of “topics.” The topic distribution of a document, we assumed that $K$ is the number of topics.

Above all, the key idea behind topic modeling is that documents show multiple topics, and therefore the key question of topic modeling is how to discover a topic distribution over each document and a word distribution over each topic, which represent an $N \times K$ matrix and a $K \times V$ matrix, respectively.

<table>
<thead>
<tr>
<th>Topics</th>
<th>Library</th>
<th>Publishing</th>
<th>Computation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>Shelves</td>
<td>Editorial</td>
<td>Computer</td>
</tr>
<tr>
<td></td>
<td>Cataloguing</td>
<td>Printing</td>
<td>Models</td>
</tr>
<tr>
<td></td>
<td>Book-ends</td>
<td>Indexing</td>
<td>Algorithm</td>
</tr>
<tr>
<td></td>
<td>Spine labels</td>
<td>Illustrations</td>
<td>Data</td>
</tr>
<tr>
<td></td>
<td>Referencing</td>
<td>Book Design</td>
<td>Information</td>
</tr>
</tbody>
</table>

Table 1: The top 5 most frequent words from 3 topics

The application of topic modeling into library electronic books immediately convert every word in the document into a “bag of words”, where probabilities are assigned to each of them, and assigned to the $K^{th}$ topic, thereby giving the user an idea of the several topics in a book. This will reduce time wastage searching through tons of pages, and perhaps overlooking books with hidden topics which the book title does not capture.

**LITERATURE REVIEW**

**Integrated library management system**

A Library Management System has been designed to automate, manage and look after the overall processing of a library, especially in ODL institutions. Efforts have been made to continually improve on library management systems, such as application login through smart
cards, RFID enabled smart library for cataloguing, circulation of materials, centralised database, user identification through their smart cards, theft detection statistics and reporting web based module etc (BGIL, 2017). More specifically, the aim is to simplify library process and in turn save time and cost.

The automated library system (ALS) has undergone significant changes since its inception in the 1970s. These changes are reflected in the conceptual differences between the ALS and the integrated library system (ILS) (Kinner, 2009). It was observed by Uzomba, Oyebola and Izuchukwu (2015) how the importance of integrated systems in library activities such as cataloguing, circulation, acquisition and serials management, etc is no longer debatable as libraries all over the world have realised the need to move from their manual practices into integrated systems and networked operations. An integrated library system can be such a robust enterprise resource management system that can continually adapt and fulfil the requirements and needs of patrons. According to Müller (2011), “In choosing ILS software, libraries must base their decision not only on the performance and efficiency of the system, but also on its fundamental flexibility to readily adapt to the future demands and needs of their patrons”. Hence the need to consciously continue to improve on these systems.

In integrating more features, its important to maintain standards, as opined by Mandal and Das (2013) that the widespread use of Integrated Library Systems (ILS), global communications via the Internet, and growing numbers of digital library initiatives have made the need for compliance with standards more critical than ever. That is, implementing information products and systems that support standards should at least ensure that library systems be able to: more easily adopt new technologies, such as, topic modelling.

**TOPIC MODELING**

A topic model, according to Liu et al (2016) is a kind of a probabilistic generative model that has been used widely in the field of computer science with a specific focus on text mining and information retrieval in recent years. The origin of a topic model is latent semantic indexing (LSI) (Deerwester et al., 1990); it has served as the basis for the development of a topic model. Nevertheless, LSI is not a probabilistic model; therefore, it is not an authentic topic model. Based on LSI, probabilistic latent semantic analysis (PLSA) was proposed by Hofmann and is a genuine topic model (Hofmann, 2001). Published after PLSA, latent Dirichlet allocation (LDA) proposed by Blei et al. (2003) is an even more complete probabilistic generative model and is the extension of PLSA. Nowadays, there is a growing number of probabilistic models that are based on LDA via combination with particular tasks. Nonetheless, all the above mentioned topic models have initially been introduced in the text analysis community for unsupervised topic discovery in a corpus of documents.

For the purpose of this paper, the Latent Dirichlet Allocation (LDA) will be considered because its current nature above others. Its a topic modeling technique that provides a generative model that describes how the documents in a dataset were created. In this context, a dataset is a collection of D documents. But what is a document? It’s a collection of words (Reed, 2012). Also according to Blei, Latent Dirichlet Allocation (LDA) is a generative
probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words (Blei, et al., 2003).

LDA assumes the following generative process for each document \( w \) in a corpus \( D \):

1. Choose \( N \sim \text{Poisson}(\xi) \).
2. Choose \( \Theta \sim \text{Dir}(\alpha) \).
3. For each of the \( N \) words \( w_n \):
   (a) Choose a topic \( z_n \sim \text{Multinomial}(\Theta) \).
   (b) Choose a word \( w_n \) from \( p(w_n|z_n, \beta) \), a multinomial probability conditioned on the topic \( z_n \).

Several simplifying assumptions are made in this basic model. First, the dimensionality \( k \) of the Dirichlet distribution (and thus the dimensionality of the topic variable \( z \)) is assumed known and fixed. Second, the word probabilities are parameterized by a \( k \times V \) matrix \( \beta \) where \( \beta_{i,j} = p(w_j = 1 | z = 1) \), which for now we treat as a fixed quantity that is to be estimated. Finally, the Poisson assumption is not critical to anything that follows and more realistic document length distributions can be used as needed. Furthermore, note that \( N \) is independent of all the other data generating variables (\( \Theta \) and \( z \)). It is thus an ancillary variable and we will generally ignore its randomness in the subsequent developments (Blei, et al., 2003).

APPLICATIONS WHERE LDA HAS BEEN USED

1. LDA-based User-Tag Model for Automatic Image Geo-Tagging
2. Applications of LDA to Document Modeling
3. Applications of LDA to Automatic Harmonic Analysis
4. Geometric Latent Dirichlet Allocation On A Matching Graph For Large-Scale Image Datasets

METHODOLOGY

HOW LDA CAN BE INTEGRATED INTO A WEB-BASED LIBRARY MANAGEMENT SYSTEM

A Library web-based system literally manages online textual and graphical resources, which in turn is made available to users for research and of course, extract information.

These are some of the features: The program called Integrated Library System with Topic Modeling is web based and was developed using C# and asp.net as its web framework. The back end is SQL based.
System Requirements
1. Hard Disk: 1G or more
2. Memory: 512MB or more
3. Processor: At least 1GHz
4. Operating System: Any OS, as long as it has browsing features. But preferably, Windows xp or later.
5. System type: 32 or 64bit
6. Software requirements: Any browser with internet connectivity. i.e. Mozilla, Internet Explorer, Google Chrome, Opera, Safari etc.

A typical library management system is developed.

Fig. 1: LMS homepage.
Fig. 2. This showcases the various functions of the LMS

Fig 3: Searching through uploaded e-books
Fig. 4: Performing topic modeling on a selected e-book.

Note: Files are inputted, variables and the output path set. Training is done on the input document and output given for analysis. The length of time spent “learning topics” is dependent on the size of document being worked on.
Fig. 5: Result after topic modeling is performed.

Fig. 6: Output folder is created and further analysis is done.

RECOMMENDATION

Further application of topic modeling should be sought and developed as its usage has continually grown in the past few years. The Open and Distance Learning platform is such a robust, scalable educational approach where easier and faster means of information exchange is sought, which can then uplift the status of ODL in the committee of institutions.

CONCLUSION

It has been seen that this approach will not only prove productive, information-wise, but also help in time management. That is, instead of flipping through already borrowed books (or tons of pages) for hours or even days, only to find out that the content is irrelevant, topic modeling should be employed to do topic extraction for well-informed decisions.

CONTRIBUTION TO KNOWLEDGE

User can now partake and explore new ways of search on future library management systems that are useful for basic tasks such as classification, novelty detection, summarisation, similarity and relevant judgments.
REFERENCES


