CTU Journal of Innovation and Sustainable Development
ISSN 2615-9422 | e-ISSN 2815-5602

DOI:10.22144/ctujoisd.2023.040

Topic based document modeling for information filtering

Nguyen Tran Diem Hanh*
School of Engineering and Technology, Tra Vinh University, Viet Nam
**Corresponding author (diemhanh_tvu@tvu.edu.vn)*

Article info.

Received 27 Jul 2023
Revised 17 Sep 2023
Accepted 2 Oct 2023

Keywords

Information filtering, information retrieval, topic models, topic modelling

ABSTRACT

Information Filtering (IF), which has been popularly studied in recent years, is one of the areas that applies document retrieval techniques for dealing with the huge amount of information. In IF systems, modelling user's interest and filtering relevant documents are major parts of the systems. Various approaches have been proposed for modelling the first component. In this study, we utilized a topic-modelling technique, Latent Dirichlet Topic Modelling, to model user's interest for IFs. In particular, an extended model of it to represent user's interest named Latent Dirichlet Topic Modelling with high Frequency Occurrences, shorted as LDA_HF, was proposed with the intention to enhance retrieving performance of IFs. The new model was then compared to the existing methods in modelling user's interest such as BM25, pLSA, and LDA_IF over the big benchmark datasets, RCV1 and R8. The results of extensive experiments showed that the new proposed model outperformed all the state-of-the-art baseline models in user modelling such as BM25, pLSA and LDA_IF according to 4 major measurement metrics including Top20, B/P, MAP, and F1. Hence, the model LDA_HF promises one of the reliable methods of enhancing performance of IFs.

1. INTRODUCTION

Nowadays, along with the development of internet technology and the huge number of content creators online, the amount of electronic contents and the number of users is increasing significantly. This leads to the problem that users are overloaded with the large amount of information. Therefore, it becomes more important to provide searching tools that can filter out irrelevant information for users so that they can approach more relevant contents to their interests quicker and with less effort. Content searching users tend to be satisfied with provided information, which are novel, familiar, important or urgent. However, these factors are not easy to determine automatically due to the overloading data and from many different resources. Regarding interest collection, there are currently two possible ways to collect information about user's interest such as implicit and explicit methods. After the user's interest over a domain of information is determined, the process of filtering relevant information that applies a certain technique in text retrieval begins searching over the data source for delivering relevant content to the users.

Topic modelling has become popular in information filtering for the past two decades. The technique was designed to model a document collection in the form of topics with the topical words extracted from the modeled documents. Among the innovative models, Latent Dirichlet Allocation (LDA) as documented by a number of studies (Blei et al., 2003; Blei, 2012; Wang & Blei, 2011) is the most popular one, providing an explicit representation of document collection. In LDA, documents in a collection can

be represented by a number of topics and each topic is a probability distribution of words. This topic model based document-representation technique has been successfully applied to many text-mining systems as the technique was able to improve the performances of relevant information retrieving capabilities in large text collection.

In LDA, topics trained from a trained collection can be used to represent the document corpus. Depending on different systems, designers can determine different numbers of topical words serving as main contents for the filtering purpose. These determining tasks definitely affect the filtering performances of the Information Filtering Systems. If the number of chosen topical words is large, this would slow down the information filtering systems. This is because the systems have to deal with a large number of calculations for providing relevant information. On the other hand, if the number of topical words is small, this results in lesser numbers of calculations, which is beneficial for the information filtering process. However, there is a downside of this approach as the filtering performance might be a disadvantage. Hence, determining the appropriate number of topical words to reach the maximum performance of an information filtering system becomes significant.

To solve the problem, we proposed a method to decide possible numbers of topical words that can help to improve the filtering performance. The model is named as Latent Dirichlet Allocation with High Frequency Occurrences, shorted as LDA_HF. The main idea of the model is to look at the word distributions in the trained topics to decide which topical words should be used to represent the topic and which topical words are not representative enough to represent the topic. In order to determine whether the new model LDA_HF is a good method in performance enhancement of Information Filtering Systems, we compared the proposed method with the existing methods in IFs including BM25, pLSA, and LDA_IF over the two large datasets including RCV1 and R8. Through extensive experiments, we found that the new model provided higher performance results than the existing methods according to four major evaluation metrics such as Top20, BP, MAP and F1.

This article is structured as follow. The first section provides an introduction and general information about the model LDA_HF. The second section provides background studies used in the study. The third section outlines the proposed model. The experiments are presented in section 4 whilst the conclusion is provided in section 5.

2. RELATED WORK

2.1. General knowledge about Information Filtering System

An Information Filtering System (IF) is similar to the information retrieval system as it provides a means for searching relevant content to searchers. In general, a typical IF system comprises of four components including (1) data-analyzer, (2) user-model, (3) filtering component, and (4) learning component.

(1) Data-analyzer: This part is used after pieces of data items was collected from information providers.

(2) User-model: This component plays an important role in representing long-term user's needs in information. Two main methods of gathering user's interest in information have been used popularly namely, explicit or implicit methods.

(3) Filtering component: This component is significantly important in any IF systems. The main task of this component is to match user information needs collected in the user-model component to the data collection before determination of data items, which are most relevant to the users.

(4) Learning component: changes in user's interest can be considered possible for improving filtering performance. This task can be carried out using this component. When user's interest changes, the filtering component should be able to detect and change accordingly for providing the users with relevant contents.

Among those mentioned components, the user model component and filtering are the most important ones, which determine the success of the retrieving performance. In fact, these components were commonly studied as below.

User model component: User interest component is important for IFs to represent how users show their preferences over the modeled collection. This component is significant because it helps to deliver relevant contents according to user's preferences. Currently the two main approaches in users' interest acquisition have been widely utilized including explicit and implicit approach. Different IF systems, different approaches in acquiring knowledge of users' preferences are utilized. Firstly, explicitly obtaining information from users is used in some

previous works (Morita & Shinoda, 1994; Yan & Garcia-Molina, 1999; Valdiviezo-Diaz et al., 2019). The major principle of the approach is that users' interest can be gathered by directly asking for users' preferences over some domains of information, areas of interest or relevant parameters. Some parameters in the process of preference collection can be documented in a form on which users were asked to fill in their preferences. Users can either show their preferences over a provided information by selecting predefined terms or show their interest level over chosen terms. In contrast, the implicit approach obtain users' preferences in information through their behaviors when making frequent transactions such as buying an item, browsing a website, searching for a piece of information, or mouse movements. Similarly, the habit of users in spending time over some items can be a good parameter to infer user preferences as suggested by (Konstan et al., 1997; Thomas & Fischer, 1996; Hu et al., 2008; Lee et al., 2008).

Filtering Component: The filtering component of an IF system carries out the search within a collection for relevant documents to an incoming document which represent user's interest. This component is the main part of any information filtering systems as it considerably affects the success of the system overall. Currently, statistical-approach and knowledge-based approach are two main techniques in filtering relevant contents. In the former approach, user models are modeled as user profiles with a weighted vector of index terms. Similarity comparison is the popular method for searching relevant contents. For example, the method of correlation and cosine measure can be used to calculate how similar an incoming document to a document in the collection. In addition, there are studies applying the ranking relevance method by using term frequency in the relevant or non-relevant documents. Another statistical approach in measuring similarity is LSI, shorted for Latent Semantic Indexing. LSI is a method, which captures the latent structure by using techniques from machine learning, can be used to retrieve relevant information. One of the noticeable features of LSI is that it can be used for finding semantic relations among terms that represent data items as described by Foltz, (1990). Similarly, Naïve Bayes (NB) classifier is popularly used in email classification systems to determine whether an incoming email is a spam email or not as provided in (Androutsopoulos et al., 2000; Lai, 2007; Sahami et al., 1998).

2.2. BM25

In BM25 as reported in (Robertson et al., 2004), terms occurring in documents of a corpus can be used to represent that corpus. This is one of the first conventional methods in representing document collections. This method was reported to be successful in some text-mining systems. In terms of techniques, the weight of a term *t* belonging to the documents is calculated as following:

$$W(t) = \frac{(tf \times (k+1))}{\left(k\left((1-b)+b \times \frac{DL}{AVDL}\right)+tf\right)} \times log\frac{(N-n+0.5)}{(n+0.5)} \quad (1)$$

Where: $W(t)$ *is* the weight of term *t*; *tf* is the frequency of term *t*; document length is *DL*; the average documents length is *AVDL*; number of documents is *N*; number of documents with term *t is n*; *k* and *b* are predefined parameters.

2.3. pLSA: Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis and introduced in Hofmann (1999), shorted as pLSA, was used for indexing documents automatically. This model has a different name as the aspect model where a statistical latent class model for factor analysis of count data is used. This statistical topic approach has been widely utilized in some text mining systems after it was introduced in 1999.

2.4. LDA: Latent Dirichlet Allocation

Among the algorithms to understand text documents, LDA as given in (Blei et al., 2003; Blei, 2012), is the best document modelling techniques, providing an explicit method for document representations. The main principle of this model is to use high frequent words in the modeled documents to represent the collection in the form of topics. In other words, a topic comprises of a number of words from the modeled documents, which can represent the modeled corpus. As reported in (Blei, 2012), LDA can be used to discover main themes of unstructured documents. In general, generating topics for a corpus can be presented as follows:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\prod_{i=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \quad (2)$$

According to Eq.(2), the major parameters of topic model are α,β and θ. Expectation Maximization (EM) in Hofmann (1999) and Gibb sampling are two popular methods in estimating

those parameters. Expectation Maximization provided by Hofmann (1999) was used to predict directly the parameters. Similarly, Gibb sampling is used for estimating the parameters, sharing common features with Markov Chain Monte Carlo.

3. THE PROPOSED MODEL

3.1. Latent Dirichlet Allocation with high frequency occurrences, LDA_HF

Definition 1: Average Distribution: Let call $W = \{w_1, w_2, \dots, w_n\}$ be the set of topical words of topic z trained over the training collection D. The probability of a topical word $w_i$to the topic z is denoted as $pr(w_i|z)$. Then the average distribution of all the topical words in W is denoted as $AvgPr(W|z)$ and calculated below.

$$\mathbf{AvgPr(W|z)} = \frac{\mathbf{1}}{\mathbf{n}} \times \sum_{\mathbf{i=1, w_i \in W}}^{\mathbf{n}} \mathbf{pr(w_i|z)} \quad (3)$$

Definition 2: High Frequent Topic Words: Let call $W = \{w_1, w_2, \dots, w_n\}$ be the set of topical words belonging to topic z trained over the training collection *D* using LDA topic modelling. Let call *X* with $X = \{w_1, w_2, \dots, w_m\}, m \leq n$ be a set of high frequent topic words in the model LDA_HF, $X \subseteq W$. A topical word in LDA_HF, $w_i \in X$ must satisfy two following conditions: (1) w_i belongs to the set of topical words trained by LDA, $w_i \in W$ and (2) the probability of the topical word w_i must be greater than average distribution of all the topical words, $pr(w_i|z) > AvgPr(W|z)$, $AvgPr(W|z)$ is measured using Eq.(3).

The purpose of high frequent words from a topic is to select topical words, which occur frequently in the topic to represent that topic. This helps to avoid picking topical words with low occurrences and hence not very representative for the topic representation.

3.2. Document ranking based on high frequency occurrences

This section presents how to calculate relevance of an incoming document *d* to the modeled corpus. It is important to note that we use term-based representation to represent the users' interest component in this work. It is also essential to emphasize that in the topic modelling technique, LDA, topic distribution provides information about how much the topic contribute to the modeled collection; distribution of words in the modeled collection represents the collection. In this work the relevance of the document *d* to the collection of documents is calculated using the document significance to the topic and the topic significance to the modeled corpus. Following is the method how to measure the relevance of an incoming document *d* based on the significance of high frequent words in that document.

Measure topic-word significance to the topic:

The significance of a topical word w_i in document *d* is denoted as $sig(w_i|z)$ and defined as below.

$$sig(w_i|z) = m_i \times Pr(w_i|z) \quad (4)$$

$$\text{Where } m_i = Pr(w_i|z) \,/\, Avg\text{Pr}(W|z)$$

$Pr(w_i|z)$ is the probability of the topical word w_i in topic *z*, $Avg\text{Pr}\,(W|z)$ is the average distribution of all topical words W of the topic z and $m_i > 1$ is selected to represent the topic in this work. This means that these topical words probabilities are larger than the average of probabilities.

Measure topic significance to the document:

Given a trained topic z_j, let$sig(z_j|d)$be the significance of the topic z_j to the document *d.* $sig(z_j|d)$ is calculated as below.

$$sig(z_j|d) = \sum_{i=1, w_i \in d, w_i \in z_j}^{n} sig(w_i|d, z_j) \quad (5)$$

where *n* is the number of selected topical words in the topic z_j; $sig(w_i|d, z_j)$ is the significance of topical word w_i to the corresponding topic.

Measure Document relevance:

For a new incoming document *d*, the document relevance score between the document *d* over the trained corpus *D* with *v* topics is measured using significance of the topic to the document and the significance of that topic to the modeled corpus as the following equation.

$$rank(d|D) = \sum_{j=1}^{v} sig(z_j, d) \times V_{j,D} \quad (6)$$

where $V_{j,D}$ is the significance of the topic z_j to the collection D, $sig(z_j, d)$ is measured using Eq.(5) and v is the number of topics trained over the collection D.

3.3. Performance evaluation

In document retrieval systems, accuracy of returning relevant documents is important. More specifically, Precision and Recall are popularly used in majority of retrieval systems. Precision determines how well the system rejects the

irrelevant documents. Recall measures how well the system is retrieving the relevant documents to a given query.

Precision is defined as the number of true positives (T_p) over the number of true positives and the number of false positives (F_p).

$$Precision = \frac{T_p}{T_p + F_p} \quad (7)$$

Recall is defined as the number of true positives (T_p) over the number of true positives plus the number of false negative (F_n).

$$Recall = \frac{T_p}{T_p + F_n} \quad (8)$$

Top-K score: *Top−K* score determines the relevance of *K* first retrieved documents. For instance, *Top−10* is the performances of the system in retrieving the first 10 documents.

Mean Average Precision *MAP*: This score determines the effectiveness of the ranking algorithm.

Break-even Point (*b/p*): This metric is used to determine the effectiveness of the filtering system. In particular, this score illustrates the points where precision and recall are equal.

F measure ($F_{\beta=1}$) The score $F1$ measures the relationship between *Recall* and *Precision*. This measurement score is the best merit among the previous measurement metrics.

$$F_\alpha = \frac{Precision \times Recall}{(1-\alpha)\, Precision + \alpha\, Recall} \quad (9)$$

Normally, $\alpha = 0.5$ is often used:

$$F_\alpha = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (10)$$

This harmonic merit mean emphasizes the importance of small values.

4. EXPERIMENTS

The following experiments were designed for proving the proposed model LDA_HF in information filtering systems. There were two hypotheses to be considered. The first hypothesis is that topic model contributes significantly to the filtering performances of IF systems. The second hypothesis is given that using high frequent topical words can help to improve the searching space for user's interest representation. This section provides detailed information about datasets, baseline models, methods to evaluate performances of the systems and experimental results.

4.1. Datasets

The Reuter Corpus Volume 1 (RCV1) datasets as given in (Lewis et al.,2004) includes articles gathered by Reuters. This datasets covers a range of domains of information and consists of the number of documents of 806.791 stories. More specifically, the datasets consists of 100 collections and is split into two main sub-datasets including the training data and the testing data. Among 100 collections, 50 collections were firstly evaluated by humans and the other collections were created by artificially incorporating the remained corpuses together. In this work, those 50 evaluated collections were used for the experiments. Datasets R8 is widely used in text retrieving systems. The data was gathered and labeled in the duration of developing the CONSTRUE text categorization systems as provided in Debole & Sebastiani (2005).

4.2. Baseline models

These following baseline experiments were carried out for measuring the effectiveness of the proposed model in IFs against the existing methods. Details of the existing models are shortly described as below.

BM25: BM25 as provided in (Robertson et al., 2004) is the model in document representation.

pLSA: This topic model used topical words to represent users' interest as provided in Hofmann (1999, 2017).

LDA_IF: This model used trained topics using LDA method to represent user's interest. The number of topical words are 20 for all the trained topics for a collection. Readers might refer to the work in Blei et al., (2003), for more detailed information.

4.3. Evaluation measurement

In the experiments, performances of the models are measured using four main evaluation metrics such as Top-20, Mean Average Precision (*MAP*), *b/p* and F_1.

4.4. Experimental results

In this section, we would like to investigate the contributions of modelling users' interest with high frequent topical words with the intention to enhance filtering performances of IFs. Experiments were conducted with prior baseline methods including BM25, pLSA and LDA_IF and the new proposed model in IFs. These experiments were carried out over 50 collections of datasets Reuter RCV1 and 8

collections in the datasets R8. For demonstration, we list all topical words in the collection 34 in datasets Reuter RCV1 trained using both models including LDA_IF and LDA_HF. In the LDA_IF model, we used topics with 20 words with high frequency to represent the collection. Table 2 displays all topical words trained over the examined collection by applying the model LDA_IF as below. Similarly, Table 1 illustrates topical words trained using the model LDA_HF in the same collection of the datasets. It is obvious that the number of topical words in the model LDA_HF is less than that of topical words in model LDA_IF because the former model chose words with high frequency to represent user's interest.

Table 1. Some topics as a result of training over the collection using the model LDA_HF

ID	Topics trained using LDA_HF
z_1	London carnival rates largest standard city holiday hill dealers complaints notting chancellor street fair Caribbean major general
z_2	violence UK staff retail retailers robbery training shop small videos train big stores wales combat pounds demand reported increasing shops
z_3	percent year increase million June recorded years number interest police incidents months thresher Tuesday
z_4	weapons handguns newspapers partys giants
z_5	ban
z_6	crime violent crimes annual risen government make cent reported fall statistics England
z_7	home liberal biggest
z_8	democrats bands
z_9	British trading independent office companies show affairs
z_{10}	company conference national minister trouble evening hours group

Table 2. Some topics as a result of training over the collection using the model LDA_IF

ID	Topics trained using LDA_IF
z_1	London carnival rates largest standard city holiday hill dealers complaints notting chancellor street fair Caribbean major general *areas multicultural ahead*
z_2	violence uk staff retail retailers robbery training shop small videos train big stores wales combat pounds demand reported increasing shops
z_3	percent year increase million June recorded years number interest police incidents months thresher Tuesday *owners tenterhooks prevention result rape lottcn*
z_4	weapons handguns newspapers partys giants *record festival abuse sale waiting prevent called order issue doubling di use west affected breath nature*
z_5	ban *police acquisitions rare live action injured subject employee chemist alex ministry threatened acquired Sunday steel event part kirkhope year*
z_6	crime violent crimes annual risen government make cent reported fall statistics England *firms rainshowers making related assaults muggings gaiety noise*
z_7	home liberal biggest *leading prowl Wednesday leave month carrying situations lottery theft country attract shot backed board visitors potentially year*
z_8	democrats bands *offer suffer stabbed stop scheme calls parts private society pound counter Thomson lunn weekend advises settled expected black*
z_9	British trading independent office companies show affairs *owners victims prime institute smallcalibre carlile told gun restrictions crowds massacre monitoring industry*
z_{10}	company conference national minister trouble evening hours group *innocent voted ignoring narrowly weeks compares legal airlines European jerry shareholder digest*

It is obvious that Table 2 contains a number of topical words, *written in Italic*, that are not representative to represent user's interest by the new model LDA_HF because of frequently occurring condition. Although these topical words are significant to the modeled collection, they still not very important to represent it in comparison to other remained topical words.

Table 3 displays the experimental results for datasets RCV1 and Table 4 present the filtering performances over the datasets R8. The percentage of change *%change* illustrates the difference between the new proposed model and the highest result among the other models in a group. The higher value of *%change*, the better the improvement of the proposed model is.

Comparison with the existing methods of user's interest representation. As displayed in Table 3, the big improvement was in Top-K score. Noticeably, LDA_HF obtained 0.464 in Top-20 whilst the scores of all term-based methods (i.e., BM25, pLSA) were 0.445 and 0.345 respectively, which changed the improvement up to 4.27%. The second big increase between the proposed model and the model LDA_IF is in MAP score, with 5.39%. In F1 score, it was 0.414 in LDA_IF and increased to 0.433 in LDA_HF, which made the improvement of the new model over the best model in that group to 4.59%.

Table 3. Comparisons among the methods for datasets RCV1

Methods	Top-20	B/P	MAP	F1
LDA_HF	**0.464**	**0.417**	**0.430**	**0.433**
LDA_IF	0.433	0.399	0.408	0.414
%Change	*7.16%*	*4.50%*	*5.39%*	*4.59%*
pLSA	0.445	0.383	0.403	0.413
BM25	0.345	0.337	0.330	0.359
%Change	*4.27%*	*8.88%*	*6.7%*	*4.84%*

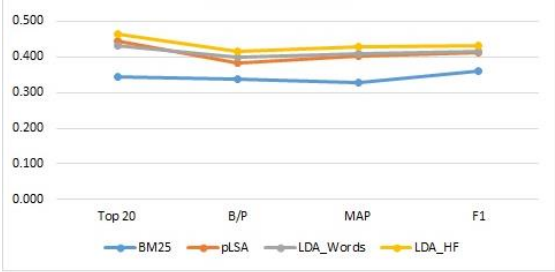

Figure 1. Performance among methods over datasets RCV1

Figure 1 shows the performances of different methods in IFs as provided in the baseline experiment section. It is obvious from the figure that the performance of the model LDA_HF is higher than other baseline models. Hence, the model of choosing frequently occuring topical words in LDA_HF has proven its enhancement in datasets RCV1.

Experiments over the datasets R8 showed a larger gap between the new model LDA_HF and the existing models in terms of Information Filtering performance. Table 4 represent the results of experiments over the datasets. Figure 2 visualized the results in line chart so that you can see the results

Table 4. Comparisons among the methods for datasets R8

Methods	Top-20	B/P	MAP	F1
LDA_HF	**0.550**	**0.561**	**0.453**	**0.440**
LDA_IF	0.506	0.371	0.362	0.354
%Change	*8.70 %*	*51.21%*	*25.14%*	*24.29%*
pLSA	0.456	0.354	0.339	0.337
BM25	0.419	0.361	0.340	0.346
%Change	*20.61%*	*55.40%*	*33.24%*	*27.17%*

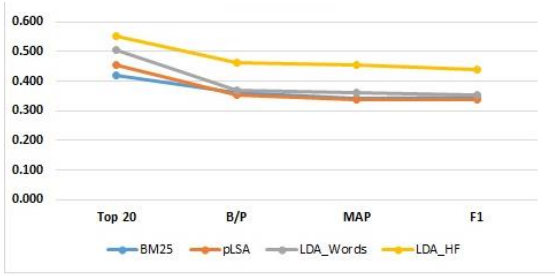

Figure 2. Performance among methods over datasets R8

In datasets R8, there was a large improvement of changes in Top-20. While it was 0.55 for LDA_HF, it was 0.456 in model pLSA, the highest performance in term-based representation. This made the change increased to 20.61%. In comparison to model LDA_IF, the change of Top-20 was 8.70% against model LDA_IF. As can be seen in datasets R8, the highest improvement is in B/P score, with 51.21%. Obviously, it was 0.371 in the model LDA_IF while it is 0.561 in LDA_HF. Similarly, the model LDA_HF was about 0.453 in MAP score whilst it was 0.362 in model LDA_IF, which increased the change to 25.14%.

According to Figure 2, the experimental results were very similar among three models BM25, pLSA, and LDA_IF. However, the results of the model LDA_HF was very different from the base line models. In other words, the proposed model showed a considerable change in retrieving relevant documents in datasets R8.

5. CONCLUSION

In conclusion, this work has provided some general concepts about information retrieval and its

importance in providing relevant contents in a huge amount of information. In particular, this work has provided a reliable method to model user's interest called LDA_HF in IFs using the word distributions in the examined topic. Extensive experiments using the new model and existing models such as BM25, pLSA, and LDA_IF in Information Filtering Systems were conducted over the two datasets including RCV1 and R8. We found that the proposed model outperformed the baseline models according to four evaluation metrics. These results have proven that the new model provides an effective model in delivering relevant contents to information searching users.

ACKNOWLEDGMENT

Thank you for your time reading this article. Should you have any comments, please contact to the author via the email provided above.

REFERENCES

Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., & Spyropoulos, C. D. (2000). An experimental comparison of Naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 160–167).

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77. doi: 10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 993-1022.

Debole, F., & Sebastiani, F. (2005). An analysis of the relative hardness of Reuters 21578 subsets. *Journal of the American Society for Information Science and technology*, *56*(6), 584-596.

Foltz, P. W. (1990). Using latent semantic indexing for information filtering. In *ACM sigois bulletin* (Vol. 11, pp. 40–47).

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (p. 50-57). ACM.

Hofmann, T. (2017). Probabilistic latent semantic indexing. In *ACM SIGIR forum* (Vol. 51, p. 211-218). ACM.

Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *2008 eighth IEEE international conference on data mining* (pp. 263–272). Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). Grouplens: applying collaborative filtering to usenet news. *Communications of the ACM*, *40*(3), 77–87.

Lai, C.-C. (2007). An empirical study of three machine learning methods for spam filtering. *Knowledge-Based Systems*, *20*(3), 249–254.

Lee, T. Q., Park, Y., & Park, Y.-T. (2008). A time-based approach to effective recommender systems using implicit feedback. *Expert systems with applications*, *34*(4), 3055–3062.

Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, *5*(Apr), 361–397.

Manning, C. D., Raghavan, P., & Schu¨tze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Morita, M., & Shinoda, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. In *SIGIR'94* (pp. 272–281).

Robertson, S., Zaragoza, H., & Taylor, M. (2004). Simple bm25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on information and knowledge management* (pp. 42–49).

Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. In *Learning for text categorization: Papers from the 1998 workshop* (Vol. 62, pp. 98–105).

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1–47.

Thomas, C. G., & Fischer, G. (1996). Using agents to improve the usability and usefulness of the world-wide web. In *Fifth international conference on user modeling* (pp. 5–12).

Valdiviezo-Diaz, P., Ortega, F., Cobos, E., & Lara-Cabrera, R. (2019). A collaborative filtering approach based on Na¨ıve Bayes classifier. *IEEE Access*, *7*, 108581–108592.

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 448–456).

Yan, T. W., & Garcia-Molina, H. (1999). The sift information dissemination system. *ACM Transactions on Database Systems (TODS)*, *24*(4), 529–565.