Automated Chat Transcript Analysis Using Topic Modeling for Library Reference Services

Xiaoju Chen*
Carnegie Mellon University
Pittsburgh, PA, USA
xjuliechen@cmu.edu

Huajin Wang*
Carnegie Mellon University
Pittsburgh, PA, USA
huajinw@cmu.edu

* These authors contributed equally

ABSTRACT

Chat reference service has been used in academic libraries to more efficiently serve patrons in the digital age. Identifying question topics on chat can help librarians understand patrons' needs and improve reference services. Researchers have used qualitative methods to understand question types in chat records; however, these methods are inefficient to analyze large chat datasets. Here, we conducted a novel research using Latent Dirichlet Allocation (LDA) topic modeling to automatically extract topics from chat transcripts generated in 5 years from a large university library. With little human intervention, the model identified major topics based on statistical distributions of terms-document relationships in chat transcripts. We also applied VOSviewer to analyze the same dataset and found consistent results. From these results, we found that the most prominent chat topics were about accessing various library resources. This finding can help libraries allocate resources, design educational materials, and provide trainings for future librarians.

KEYWORDS

Text mining; topic modeling; VOSviewer; chat reference; Q&A;

ASIS&T THESAURUS

Topic models; Needs assessment; Academic libraries

INTRODUCTION

Chat virtual reference services have been used by academic libraries to provide real-time online reference services (Lee, 2004). Understanding questions frequently asked through chat can help librarians better understand patrons' needs and provide more timely and effective answers. To analyze question types and topics, researchers have traditionally used qualitative research methods to examine chat records, usually with a relatively small dataset collected. Diamond and Pease studied 450 reference transactions during a two-year period and concluded that standard reference questions were most frequently asked (Diamond & Pease, 2001). Lee measured questions asked through chat for a six-month period and found that accessing database and electronic resources and administrative questions represented over 60% of the questions asked (Lee, 2004). Recently, with the development of new technology, researchers began to use more advanced quantitative analysis methods to analyze much bigger datasets. For example, in a mix-method study to understand learning mechanisms on the chat platform used in an academic library (Schiller, 2016), a line-by-line open coding on 1% of the data were first conducted, before applying a text mining tool (QDA miner) to analyze the remaining records. These qualitative and mixed method studies are labor intensive and cannot keep up with the growing amount of data collected on electronic reference platforms nowadays. It is therefore necessary to explore automated methods using unsupervised probabilistic machine learning models that is more scalable.

In recent years, topic modeling methods have matured as useful tools to extract topics from documents, without the need to label the datasets or having predefined taxonomies. Latent Dirichlet Allocation (LDA) has been discussed as one of the best suited methods for finding topics and types in text document (Linares-Vásquez, Dit, & Poshyvanyk, 2013)(Fu, 2017), and has been used to analyze topics on web-based question and answer (Q&A) platforms such as StackExchange and StackOverflow. Barua et al used LDA to analyze the main topics in the software developer discussions (Barua, Thomas, & Hassan, 2014). Allamanis and Sutton used a similar method to associate programming concepts with particular types of questions asked through StackOverflow (Allamanis & Sutton, 2013). Fu performed a LDA to understand music topics covered by questions and answers on StackExchange (Fu, 2017). These studies provided insights into how LDA can be used to analyze Q&A type of records and help to discover main topics and questions types. However, it has never been applied to understanding library reference Q&As.

Apart from topic modeling methods, network analysis tools such as VOSviewer are widely used to analyze topics and their relationships in large-size text records. VOSviewer was developed by Leiden University for analyzing bibliometric network data (van Eck & Waltman, 2011). It can cluster most important terms in a text input by finding noun phrases and

82nd Annual Meeting of the Association for Information Science & Technology
Melbourne, Australia | 19 – 23 October, 2019

their frequencies and co-occurrences. Due to its ability to show key phrases and their relationships, researchers have been using VOSviewer as both text mining and visualization tool to identify topics in scholarly outputs (Park & Nagy, 2018) (Mahieu, van Eck, van Putten, & van den Hoven, 2018) (Demeter, Szász, & Kő, 2019).

In this study, we built an LDA model to identify major topics occurred in library reference Q&As using chat transcripts in the last 5 years. The same dataset was also used in parallel to generate a term map using VOSviewer. Results generated from both methods were compared to provide further quality assurance for topics extracted from chat records.

METHODOLOGY

Data collection and preprocessing

The chat service offered by the university libraries has been provided using a third-party platform, LibraryH3lp. To generate the raw dataset, we exported all chat transcripts from January 1, 2013 to December 31, 2018 as CSV format. Each chat interaction often includes multiple rounds of questions and answers, and is considered as one record. After filtering out empty records and system generated offline messages, there were 5610 records with approximately 1.3 million words for our analysis. We used home-made Python (version 3.5.4) scripts to remove system-generated information and parse text. To further prepare the dataset for topic modeling, we used the Python NLTK Toolkit (version 3.2.5) to remove stopwords, tokenize, and lemmatize. To customize our model to the library services setting, we compiled a list of customized stopwords in addition to the standard English stopwords built-in with the Toolkit, e.g. greeting words.

Topic modeling with Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is one of the most common probabilistic topic models used to automatically discover topics hidden in a collection of documents. It assumes that a document has a number of topics and these topics can be characterized by a distribution over words. By calculating document-topic and topic-word distributions using Dirichlet distribution as a prior, the model automatically generates topics and a set of keywords associated with each topic based on their probability distributions in the given text documents. The meaning of each topic is interpreted and determined by humans. Here, we implemented the LDA model using Python Gensim library (version 3.4.0) (Řehůřek & Sojka, 2010), which is based on the online variational Bayes algorithm of LDA implementation (Hoffman, Bach, & Blei, 2010). To train the model, we ran the LDA model for 20 passes, with random_state sets to 50 to enable reproducibility.

To determine the best number of topics K, we ran the LDA algorithm with K ranging from 2-20, and evaluated the quality of topics by calculating perplexity and coherence scores. Cv metric was used to compute coherence scores as it was reported to perform the best over many benchmark datasets (Röder et al. 2015). Furthermore, we evaluated the quality of topics by visualization of topic clusters using pyLDAvis library (version 2.1.2), as well as by manually inspecting topics under each K. Based on these methods, we set the model to K=8, which produced a perplexity score of -7.38 and coherence score (Cv metric) of 0.44. Labels of the topics were presented as serial numbers when generated by the LDA model, and later interpreted by human experience based on top keywords present in each topic.

Constructing distance-based maps with VOSviewer

One of the difficulties in using topic modeling for text mining is model evaluation. To gain further confidence in the topics generated from the LDA model, we ran the same preprocessed dataset with VOSviewer (version 1.6.10) to generate distance-based term maps. We used full-counting method (total numbers of occurrences in all documents) to analyze the 60% most relevant terms occurred at least 10 times within the entire dataset. Association strength was used for normalizing the strength of links between items, and resolution value was set to 1.2.

RESULTS AND DISCUSSION

Once model parameters were fixed as described in the METHODOLOGY session, we ran the LDA model multiple times to observe the stability of topics produced. Output from the model was presented as keywords for each topic and the weightage (importance) of each keyword. The content of each topic was determined by reference librarians based on the interpretation of the keywords in each topic. Results derived from a representative run is shown in Table 1, showing top 10 keywords in each of the 8 topics identified by our LDA model. Topics were sorted by percentage of tokens of a given topic, with T1 accounting for the most percentage (22.3%) of total tokens (Figure 1, right). In each topic, keywords were arranged in descending order according to their weights. By running the model multiple times, we observed that most topics are very stable, including T1 (***physical book access***), T2 (***journal article access***), T3 (***off-campus access***), T4 (***interlibrary loan***), T6 (***guest access***), and T7 (***thesis and dissertation***). T5 represents a typical topic (***specialized reference***), in which specific subjects often co-occur with names of subject librarians in most runs, but in this specific run, those keywords were ranked lower. T8 (***http link to catalog item***) is the only topic that did not appear in every run.

To help intuitively assess the topics, we visualized the model output using Python pyLDAvis library (Figure 1). The result shows that for some topics, there are overlaps among each other, because terms can belong to more than one topic, while others are well separated from each other (Figure 1, left). It is worth noting, however, the visualization is a mapping of the multidimensional data into a 2-D plane through multidimensional scaling; the actual separation in higher dimensions are likely to be better.

To further verify the quality of the LDA model and the results, we analyzed the same preprocessed text with VOSviewer. VOSviewer uses a different text mining algorithm to analyze and visualize topics. Results generated by VOSviewer (Figure 2) show that the largest topic was ***physical book*** (dark blue circles), example terms include “book”, “circulation desk”, “hunt library”. Additionally, ***journal article access*** (light green circles; key words: “e-journal”, “volumn”, “issue”), ***off-campus access*** (green circles; key words: “off-campus”, “vpn”, “proxy”), and ***interlibrary loan*** (yellow circles; key words: “request”, “interlibrary loan”, “business day”) were also hot topics. These hot topics identified by VOSviewer were consistent with the those identified by the LDA model. VOSviewer also identified an additional topic, ***library account*** (sky blue circles; key words: “number”, “account”, “login”), that was not shown in one of the eight LDA topics. However, based on the keywords, the ***library account*** topic can be seen as an access question associate with electronic resource, which was captured in the LDA model.

ID	Topic	Keywords (top 10)
T1	Physical book access	book, library, check, available, see, help, hunt, get, hold, borrow
T2	Journal article access	article, journal, access, search, database, link, see, find, help, looking
T3	Off-campus access	access, library, id, vpn, link, campus, try, get, log, using
T4	Interlibrary loan	request, loan, get, ill, interlibrary, illiad, need, article, email, library
T5	Specialized reference	librarian, help, email, contact, information, find, question, know, liaison, looking
T6	Guest access	library, access, student, need, university, help, get, use, know, public
T7	Thesis and dissertation	dissertation, thesis, check, copy, help, online, find, title, looking, library
T8	http link to catalog item	citation, m, copy, author, style, volume, v, carnegie, j, vol

Table 1. The 8 topics and top keywords associated with each topic discovered by the LDA model. Names of the topics are generated based on human interpretation.

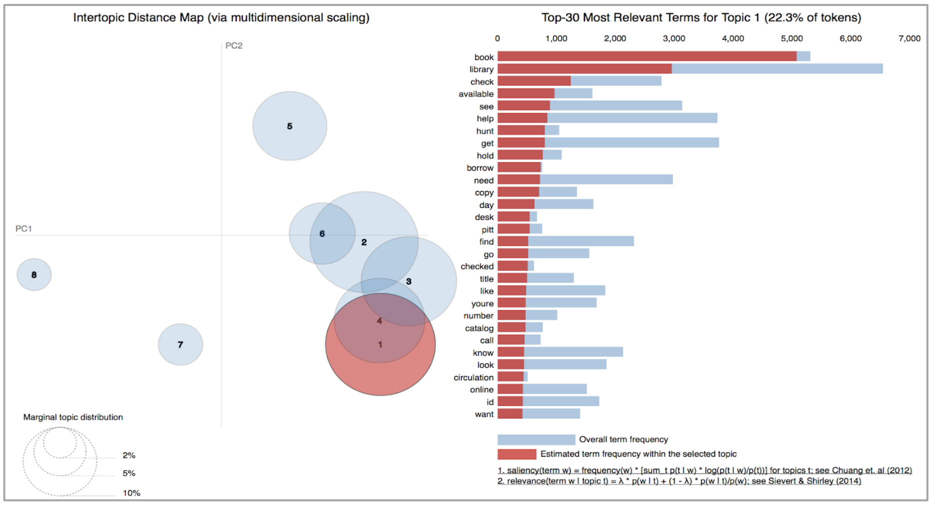

Figure 1. Screenshot of interactive visualization output from pyLDAvis. Left: distance map created based on keywords occurrence. Each cluster represents a topic generated by the LDA model. The index number of each cluster corresponds to the topic ID in Table 1. Right: distribution of the top 30 most relevant terms among topics. Red: the current topic; blue: other topics.

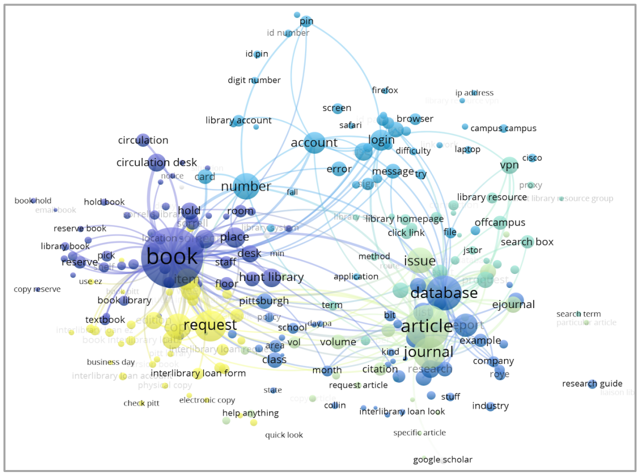

Figure 2. Screenshot of a representative distance map generated by VOSviewer, using the same preprocessed dataset used to build the LDA model. Colors show different clusters.

Both LDA and VOSviewer results show that ***physical book access*** was the most mentioned topic in chat conversations (T1), followed by various ***access*** topics (T2, T3 T4, and T6). These topics also had more interconnections, reflected by the visualization in VOSviewer: terms in these hot topics such as ***book*** (dark blue) and ***interlibrary loan*** (yellow) were closely connected. This finding was consistent with the practice in real-life library services. Patrons often use the chat service for a fast and straightforward solution, for example, when they have problems accessing an electronic item. These problems are often caused by 1) the patron was not using campus network so they cannot use library resources, or 2) the library does not have subscription to the item. Therefore, ***off-campus*** and ***interlibrary loan*** were frequently mentioned topics in chat and these topics interconnect with ***book access*** and ***journal access***. Apart from the ***access*** topics, we found three isolated topics (T5, T7 and T8). In T5 (***specialized reference***), LDA keywords “librarian”, “email”, and “liaison” showed that patrons were referred to subject librarians for better assistance. T7 (***thesis and dissertation***) were questions related to locating thesis and dissertations published by graduate students in the university. T8 (***http link to catalog item***) included words and terms in the university library website’s URL, inked to specific items in the catalog. In the eight topics discovered by the LDA model, most of the access questions can be answered by staff and students at the circulation desk; while T5 and T7 should be answered by more experienced and/or subject librarians.

CONCLUSION

We collected electronic chat transcripts from the past 5 years in the university library and used the dataset to build an LDA topic model. The model was able to identify 8 major topics related to Q&As using the chat reference service. Findings from this study will provide a data-driven approach for academic libraries to identify patrons’ needs regarding the use of library resources, and help libraries to make informed decisions on how to allocate resources, design educational materials for patrons, and provide training for future librarians.

In the future, we will apply our model to analyze evolving trend in Q&As over the years, and extend this model to analyzing other electronic reference questions, such as emails and web forms.

REFERENCES

Allamanis, M., & Sutton, C. (2013). Why, when, and what: Analyzing Stack Overflow questions by topic, type, and code. *2013 10th Working Conference on Mining Software Repositories (MSR)*, 53–56.

Barua, A., Thomas, S. W., & Hassan, A. E. (2014). What are developers talking about? An analysis of topics and trends in Stack Overflow. *Empirical Software Engineering*, *19*(3), 619–654.

Demeter, K., Szász, L., & Kő, A. (2019). A text mining based overview of inventory research in the ISIR special issues 1994–2016. *International Journal of Production Economics*, *209*, 134–146.

Diamond, W., & Pease, B. (2001). Digital reference: a case study of question types in an academic library. *Reference Services Review*, *29*(3), 210–219.

Fu, H. (2017). Categorization of musicology questions from community-based Q&A site using latent dirichlet allocation. *IConference 2017 Proceedings.*

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online Learning for Latent Dirichlet Allocation. In J. D. Lafferty,

Lee, I. J. (2004). Do Virtual Reference Librarians Dream of Digital Reference Questions?: A Qualitative and Quantitative Analysis of Email and Chat Reference. *Australian Academic & Research Libraries*, *35*(2), 95–110.

Linares-Vásquez, M., Dit, B., & Poshyvanyk, D. (2013). An exploratory analysis of mobile development issues using stack overflow. *2013 10th Working Conference on Mining Software Repositories (MSR)*, 93–96.

Mahieu, R., van Eck, N. J., van Putten, D., & van den Hoven, J. (2018). From dignity to security protocols: a scientometric analysis of digital ethics. *Ethics and Information Technology*, *20*(3), 175–187.

Park, J. Y., & Nagy, Z. (2018). Comprehensive analysis of the relationship between thermal comfort and building control research - A data-driven literature review. *Renewable and Sustainable Energy Reviews*, *82*, 2664–2679.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM ’15*, 399–408.

Rossum, G. van. (1995). *Python tutorial* (No. CS-R9526). Amsterdam: Centrum voor Wiskunde en Informatica (CWI).

Schiller, S. Z. (2016). CHAT for chat: Mediated learning in online chat virtual reference service. *Computers in Human Behavior*, *65*, 651–665.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70.

van Eck, N. J., & Waltman, L. (2011). Text mining and visualization using VOSviewer. *ArXiv:1109.2058 [Cs]*. Retrieved from http://arxiv.org/abs/1109.205