Exploring the Music Library Association Mailing List: A Text Mining Approach

Xiao Hu[1] Kahyun Choi[2] Yun Hao[2] Sally Jo Cunningham[3] Jin Ha Lee[4]
Audrey Laplante[5] David Bainbridge[3] J. Stephen Downie[2]

[1]University of Hong Kong
xiaoxhu@hku.hk

[2]University of Illinois
{ckahyu2, yunhao2,
jdownie}@illinois.edu

[3]University of Waikato
{sallyjo, davidb}
@waikato.ac.nz

[4]Univeristy of Washington
jinhalee@uw.edu

[5]Université de Montréal
audrey.laplante@umontreal.ca

ABSTRACT

Music librarians and people pursuing music librarianship have exchanged emails via the Music Library Association Mailing List (MLA-L) for decades. The list archive is an invaluable resource to discover new insights on music information retrieval from the perspective of the music librarian community. This study analyzes a corpus of 53,648 emails posted on MLA-L from 2000 to 2016 by using text mining and quantitative analysis methods. In addition to descriptive analysis, main topics of discussions and their trends over the years are identified through topic modeling. We also compare messages that stimulated discussions to those that did not. Inspection of semantic topics reveals insights complementary to previous topic analyses of other Music Information Retrieval (MIR) related resources.

1. INTRODUCTION

The Music Library Association Mailing List (MLA-L)[1] has a variety of subscribers including music librarians, MLA members, and professionals and students in music librarianship as well as musicology. The archive of this list serves as an invaluable resource for studying discussions among these people; it is the oldest and largest repository for studying issues in this profession [11].

Music librarians are an integral and important part of the larger Music Information Retrieval (MIR) community. The experiences, expertise, interests and concerns of music librarians are highly relevant to the advancement of MIR research. Similarly, MIR research can improve practices in music librarianship, ultimately enhancing the discovery of music information for diverse types of users [23]. Given the abundance of emails archived in the MLA-L, we can identify main topics discussed throughout the years which can offer insights into real-world music information interactions, particularly concerning the use and management of music information, from the perspectives of music information professionals and their clients. This study seeks to uncover the key topics of discussion related to music information needs and uses in the MLA mailing list by using text mining and quantitative analysis methods.

2. LITERATURE REVIEW

2.1 MLA-L Content Analysis

Interest in, and analysis of, the MLA-L collection is not new: nearly three decades ago when e-mail was the newest form of written communication, the list became the object of investigation (e.g., [5], [26]). Griscom [11] offered a detailed and relatively more recent account of the history and development of the MLA-L. Through qualitative content analysis, Griscom organized postings in the "E-Mail Digest" column of the MLA-L archive into nine categories as shown in Table 1.

Category	Definition
Reference questions	questions on locating songs or music work of specific topics or sources
Cataloging	extended discussions on catalogs of music library collections
Practical matters	problems unique to music libraries such as circulation and preservation of holdings
Technology	questions and opinions about adapting to technological advancements
Ethics	questions on unexpected and controversial topics such as illegal items
Copyright	questions and comments on reproduction matters and copyright laws
Circulation policies	policies and procedures posed by special formats in music libraries
Assisting colleagues	alerts on problems and peculiarities such as production errors.
MLA matters	communications from the board of directors of the association

Table 1. Main categories of MLA-L postings in [11].

While it is noteworthy that a majority of these categories were consistent with the categorization in earlier studies on the MLA-L [5], [26], analytical research on mailing lists or discussion forums of library professionals

[1] https://www.musiclibraryassoc.org/?page=mlal

should not only identify categories of messages, but also consider what specific topics were discussed [10]. More importantly, the data analyzed in the most recent endeavor of MLA-L content analysis [21] were up to year 1998 which was nearly two decades ago, and thus inspiring the updated account via current exploration of the data.

2.2 Text Analysis in MIR

Text analyses of MIR-related resources have increasingly appeared in recent years in the MIR community. Downie and Cunningham [9] contributed an early analysis of postings of music-related information requests on a music newsgroup. In their results, "locate" was identified as a predominant intended use for the requested music information. This is consistent with the category of "reference questions" shown in Table 1. In fact, the category "reference questions" was identified in all the previous studies on the MLA-L [5], [11], [26].

Music related Q&A (question & answers) websites are also a rich resource of MIR-related discussion. Bainbridge et al. [1] analyzed users' music queries in Google Answers using a grounded theory approach, revealing that bibliographic metadata were commonly included in users' queries. This leads to corroboration with the "cataloging" category of MLA-L postings [11]; bibliographic metadata such as *performer* and *title* are necessary for catalogs in music libraries. A later study by Lee [16] employed content analysis, also on Google Answer queries, to look into music information seeking behaviors. The results revealed that the "location" of a music information object (e.g., a recording) was also one of the most prominent information needs.

In addition to newsgroups and Q&A websites, publications in MIR have also been analyzed. Lee et al. [16] examined papers in the proceedings of the Conference of the International Society for Music Information Retrieval (ISMIR). Analysis of keywords in titles and abstracts of ISMIR papers identified "audio" and "classification" as the most frequent terms. Most recently, Hu et al. [13] compared the keywords in titles of ISMIR papers written by female to those by male authors, revealing the gender-based differences of topic preferences between authors.

Lyrics, music reviews, users' interpretations of music, and other types of listeners' input in social media (e.g., social tags and tweets) have also been analyzed, mainly by automated methods [24], for various tasks in MIR such as genre classification [22], mood classification [14], and subject classification [7]. In these studies, natural language processing methods are applied to convert textual data into numerical data in large scales, which are then fed into a wide range of machine learning approaches to fulfill aforementioned MIR tasks. More often than not, such texts are used in combination with audio signals to further improve performances via multimodal approaches (e.g. [14]).

2.3 Topic Modelling and Trend Analysis

Machine learning and quantitative methods have demonstrated their capability in automating the processing of email messages [6] and other text input by users. Barua et al. [2] used a prevalent statistical modeling technique called Latent Dirichlet Allocation (LDA) [3] to discover topics and their trends in Q&A websites, in order to gain insights into the wants and needs of the participants. Prior studies employing topic modelling on email messages and replies are seemingly scarce. McCallum et al. [20], being one rare case, also used the LDA in analyzing an email corpus.

Exploration of the changes in topics found in a body of messages over time—trend analysis—is particularly important due to the evolving focus in documents such as emails and queries (Blei and Lafferty, 2003 as cited in [27]). Unsupervised topic modeling as an extension to LDA is one way to generate the temporal relationships of topics [12].

Mishne and Glance [21] showed that comments made by users on weblogs could be an indicator of popularity of posts or the weblogs themselves. The mechanism of MLA-L is also in the form of postings welcoming potential replies, and thus it is reasonable to evaluate the popularity of topics based on their corresponding replies.

To bridge the gaps in previous research, this study aims to answer the following research questions: 1) What are the primary topics discussed in the MLA-L list from 2000 to 2016?; 2) How did the strength of the topics change over time?; and 3) Which topics attracted replies and which did not? Answers to these questions will help MIR researchers and practitioners understand information needs of the community and identify potential use cases of MIR tasks and applications.

3. DATA STATISTICS

The corpus used in this study consists of 53,648 emails posted on MLA-L from 2000 to 2016 by approximately 2,713 people (Figure 1). Among these emails, 33,250 (61.98%) received no replies while the other 20,398 emails (38.02%) formed 8,384 distinct online conversations (email threads) with the largest thread containing 52 emails. The average length of an email is 177.5 words (after removing reply and signature blocks; blank emails not included) with the longest email containing 1,389 words, indicating that most of the emails contain substantial content.

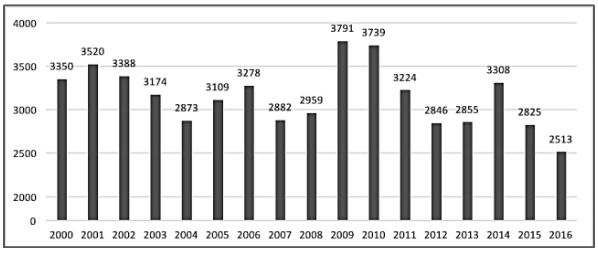
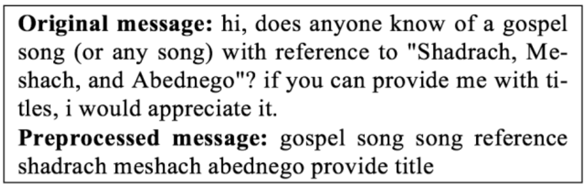


Figure 1. Number of emails in MLA-L across years.

4. PREPROCESSING

The data cleaning steps for preprocessing the large corpus include: 1) removing special formats and converting emails into plain text using OpenRefine[2]; 2) removing reply blocks (quoted emails included in reply emails); 3) detecting signature blocks by detecting variations of the sender's name using the tool Jangada[3] from the last 10 lines of the emails as suggested in [6]; 4) removing mailing list footers; and, finally 5) removing various leave-takings such as "best wishes" and "sincerely." Given the large amount of data, it is not possible to manually evaluate the accuracy of preprocessing on the entire corpus. Instead, inspection of a random sample of 100 preprocessed emails show 89% had the aforementioned noisy parts correctly removed.

The cleaned emails underwent text processing procedures for text mining purposes. Stopwords[4] that are extremely common in English were removed to increase the quality of discovered topics. In addition, words that occurred in more than 15% of the emails, such as "music" and "send", were also removed, since they had little discriminant power and thus could be regarded as domain stopwords. To enhance the readability of the topics, lemmatization[5] was applied to convert words in derivative forms into their lemmas, instead of crude stemming that usually cuts out suffixes of words. Finally, emails that were left with less than four words were eliminated as it would not be reliable to assign them to a certain topic based on too few words. Figure 2 illustrates an example email before and after preprocessing.



Original message: hi, does anyone know of a gospel song (or any song) with reference to "Shadrach, Meshach, and Abednego"? if you can provide me with titles, i would appreciate it.

Preprocessed message: gospel song song reference shadrach meshach abednego provide title

Figure 2. An example email from the MLA-L postings before and after preprocessing.

5. METHODS

5.1 Topic Modeling Setup and Labeling Procedure

We employed one of the popular topic modeling algorithms, Latent Dirichlet Allocation (LDA). It is a generative model that represents documents as probability distributions over topics, and represents topics as probability distributions over words [3]. In other words, given a set of documents, LDA can identify latent topics discussed in these documents based on word-document co-occurrences. As a probabilistic method, LDA can assign each document to a small number of topics with different probabilities. At the same time, each topic is represented by a small set of words that are highly related to this topic. Due to its superior performances compared to other topic modeling methods, LDA has been widely used to discover topics from diverse corpora such as papers, web postings, and even tweets. It is believed that LDA can also discover topics well from email content [20].

In this study, we ran the LDA implementation in the MALLET machine learning toolkit, which has been widely used in topic modeling research [19]. In LDA, the number of topics is tightly linked to the granularity of the learned topics. After changing the number of topics from 10 to 200, we report the case when the number of topics was set as 50, for a proper granularity of topics. To increase the quality of text analysis, we used bigrams (i.e., combinations of two consecutive terms) as well as unigrams (i.e., individual terms) as "words" when learning the topics [25].

The resultant topics then underwent a manual screening process, to filter out noisy, meaningless topics. This is a common practice in applying topic modeling (e.g., [12]). Those noisy topics were introduced by trivial text patterns that frequently appeared in the corpus. In our context, due to the inevitable side effect of the automatic approach to data preprocessing, which was mandated by the scale of the corpus, some signatures and forwarded message headers remained and were inputted into the topic modeling process. As a result, these signatures and senders were grouped into a noisy topic. Another example of a noisy topic is the agglomeration of words in languages other than English. Common email terms (e.g., "post", "email") and greeting/closing words also formed a topic which conveyed little meaning. Fortunately, topic modeling associates each identified topic with representative words, and thus it is convenient and reliable for researchers to examine and weed out noisy topics.

The remaining topics appear to be meaningful. To enhance readability of the topics, we manually labeled each topic with one or two phrases based on the meanings of the top 10 words and the top 50 emails associated with the topic. In particular, frequently appearing words in the subject lines of the highly ranked emails in each topic help us gain a deeper understanding of the topic.

5.2 Topic Trend Analysis

Besides identifying topics in the entire corpus, we also calculated the probability over topics given a particular year, $P(t|y)$, to examine the topical trend over time. To this end, we followed the empirical probability calculation procedure proposed in [12]. Based on the topic modeling result, we first computed a matrix $\boldsymbol{C}$ of $D\times T$ dimensions where D refers to the number of documents (emails) in our corpus and T the number of topics identified. The (d,t)-th element of the matrix $\boldsymbol{C}$ holds the number of words assigned to the t-th topics in the d-th document. From here, we can induce the probability over the topics per year y as follows:

[2] http://openrefine.org/

[3] http://www.cs.cmu.edu/~vitor/codeAndData.html

[4] https://code.google.com/archive/p/stop-words/

[5] http://morphadorner.northwestern.edu/morphadorner/

$$P(t|y) = \frac{\sum_{d=D_y} \boldsymbol{C}(d,t)}{\sum_{d=D_y} N_d}, \qquad (1)$$

where D_y is the set of documents that belong to year y, and N_d is the number of words in the d-th document. The probability $P(t|y)$ over the years can then form a time series for topic t.

In order to determine whether there is a statistically significant upward or downward trend for a topic over time, Cox Stuart trend analysis was used with a significance level of 0.05 [18]. Cox Stuart trend analysis splits a time series into two halves and counts the numbers of positive and negative differences between pairs of data points drawn from the two halves. If there are more negative differences than positive ones, then an upward trend is detected, indicating that the topic gained more attention over the years, and vice versa.

5.3 Topics with and without Replies

To answer the third research question, we compared the topics associated with emails with replies and those without. As mentioned before, in the results of LDA, each email is assigned to multiple topics with different probabilities. In this analysis, we aggregated the topics and their probabilities across all emails either with or without replies. Then we ranked the topics based on their aggregated and normalized probabilities. By comparing the top topics of the two sets of emails, we can discover which topics were more engaging and generated more discussions among the MLA community.

6. RESULTS AND DISCUSSION

6.1 Discovered Topics

Upon careful screening, 27 of the resultant topics were identified as meaningful. Table 2 presents these topics, ranked by their topic weights. For each topic, Table 2 presents the topic IDs, the labels, the topic weight (in parentheses), frequent words in the subjects of top emails associated with this topic, the top 10 words assigned to this topic in the LDA results, trend over time (upward or downward as indicated by arrows), and the trendline that plots the probability change of this topic across the years.

The weight of the topic (the Dirichlet parameter) is roughly proportional to the overall portion of the documents assigned to a given topic [19]. Therefore, topics with higher weights are more popular in the corpus, while those with lower weights rarely appear. As shown in Table 2, it is not surprising that the most important topic in our corpus is about requests and questions from patrons (i.e., clients). Unlike other topics, there is no frequent word in subject lines for this topic. A closer examination revealed that it is because each subject belonging to this topic was unique. Based on this topic's highest weight, we can infer that there must have been a wide range of requests and questions from patrons. The second highest ranked topic is musical terms whose top words include a range of music genres, indicating that music librarians not only focus on Classical music (topic #35), but also on a wide diversity of music. Other top ranked topics cover various aspects of librarianship (e.g., #6: cataloging; #14: circulation; #22: collection), and music-specific materials: scores (#30), recordings (#47), and songs (#12).

In order to compare our results to categories identified in previous studies using content analysis, the 27 meaningful topics discovered in this study were manually grouped into nine broader topic categories based on their semantic similarity: Cataloging, Reference Questions, Circulation Policies, Copyright, Audio Technology, MLA, Advertisements, Music Related Terms, and Others. Five of the categories (Reference questions, Cataloging, Copyright, Circulation Policies, MLA) were equivalent with those uncovered by Griscom [11] (c.f. Table 1), while Audio Technology appears closely related to Griscom's "Technology." The Advertisement (CD/DVD sales, Travel information, Job postings) category is novel to our findings.

Our discovered topics are also consistent with results from earlier topic analyses on MIR-related discussions. For example, some of the most frequent words in topic #31 (e.g., "bibliographic," "metadata") are exactly the same as how users of MIR systems predominantly described their needs, particularly on music-related discussion platforms [1], [9]. Similarly, Audio Technology was one of the main topics revealed in this analysis, whereas "audio" has been one of the most commonly used title terms in ISMIR research topics [16]. With the rapid growth of audio-based research in the MIR community, insights and needs on audio technology from the music librarian community can provide important real-life use cases for MIR studies. Another example is Topic #42 which was labelled "Grove music online." It echoes the trend of online access to digital music information, which is also a major theme in MIR research community. As a major research-oriented online resource serving scholars and music professionals, Grove Music Online can be used by MIR researchers for improving MIR services and applications targeting the scholarly and professional user groups.

6.2 Topic Trend Analysis

The temporal trends of these topics are reported in the "trend column" in Table 2. Results of Cox Stuart tests show that six topics had increasing trends (#20, #31, #7, #49, #24, #27.) while four had decreasing trends (#41, #14, #12, and #2). Other topics had no significant trend. It is noteworthy that the topics with decreasing trends ranked higher in Table 2 than those with increasing trends, reflecting a phenomenon that popular topics became less dominating over time while topics with lower weights started gaining popularity in recent years, resulting in diversified topics.

The topics showing downward trends are related to the traditional functions of music libraries: #41 (Patron's requests and questions), #14 (Circulation and library policy in colleges), and #12 (Song requests). These indicate

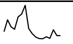
ID	Topic Label (weight)	Frequent words in Subjects	Top 10 Words	Trend	Trendline
41	Patron's request & question (0.086)	N/A	dear, copy, cw, patron, source, dear, cw, piece, score, advance, check	↓	
28	Musical terms (0.062)	Jazz, band, blues, rock music, songs	Jazz, band, record, blue, play, album, rock, live, john, sing	-	
4	MLA Event (0.054)	MLA meeting, conference, mentoring	mla, meeting, program, meet, conference, session, attend, pm, friday, time	-	
6	Cataloging (0.054)	oclc, lc, classification, cataloging, authority record, bib record	Record, title, oclc, number, catalogue, authority, catalog, add, authority record, item	-	
14	Circulation and library policy in colleges (0.046)	school, due date, except to faculty, circulate, Music Practice Room(s) in Library, students, faculty	student, collection, faculty, material, item, class, patron, circulate, score, staff	↓	
35	Classical music (0.045)	orchestra, chamber music, piano, concerto	piano, orchestra, symphony, violin, string, quartet, op, concerto, sonata, piece	-	
22	Music collection (0.039)	Music collection, catalogue, material, archive, donation	collection, manuscript, sheet, material, book, score, item, special, rare, project	-	
30	Scores, edition (0.037)	score, edition, need, actual music piece information (op, major, ..)	score, edition, publish, publisher, volume, copy, print, major, publication, complete	-	
47	Audio recording (0.034)	audio, recording, streaming, iPod, I-tunes, physical format, mp3	naxos, audio, recording, classical, stream, file, listen, digital, service, record	-	
12	Requesting songs (0.032)	Folk song for a wedding, Song in a Sopranos episode, Animal Songs, Songs about aging, Lyrics question,	song, lyric, sing, tune, word, folk, title, tina, popular, dallas	↓	
23	Journal and periodical (0.030)	journal, JSTOR, RIPM publications, ECO music journal, IIMP	journal, article, review, publish, issue, rilm, online, editor, title, publication	-	
39	Job posting (0.028)	Job opening, Job posting, position, job announcement	service, librarian, experience, collection, position, reference, application, degree, faculty, professional	-	
2	Music storage (0.028)	cd, lp, case, vinyl cd sleeves, cd box lids	cd, disc, lp, dvd, record, tape, label, case, box, booklet	↓	
20	MLA board, member (0.027)	MLA Newletter, roundtable, Note-Book, Call for new members, Board meeting, Board reports	mla, committee, member, board, association, report, chair, membership, year, annual	↑	
33	List of books (0.027)	Encylopedia, books, bibliography, Reference titles to give away, book titles	book, press, author, title, publication, york, year, isbn, publish, history	-	
0	Subject Heading, lc, genre, code (0.027)	call numbers(lc, Dewey), language code zxx, Genre heading, field, marc, aacr2, classification	subject, head, heading, term, title, instrument, score, code, form, musical	-	
29	Copyright (0.025)	Copyright question, copyright tips, copyright courses, royalties, purchasing a download, ILL(InterLibrary Loan)	copyright, copy, law, fair, public, license, domain, legal, public_domain, permission	-	
44	Journal and periodical (0.024)	Journal, issue, periodical	issue, journal, volume, spring, copy, fall, american, june, summer, july	-	
10	Conference roommate/transportation -mate solicitation (0.022)	roommates for, Shuttles to, registration	registration, conference, hotel, room, rate, register, tour, roommate, reservation, fee	-	
31	Metadata (0.020)	Music metadata, RDA, MOUG, Bibliographic Control Committee (BCC), ISBD, OCLC-MARC, music cataloging	catalogue, rda, bibliographic, metadata, marc, moug, oclc, service, indiana, access	↑	
7	Call for papers, proposals, awards, etc. (0.019)	Call for Papers, Call for Submissions, Call for Proposals, Call for Applications, Call for poster sessions, call for seminar topics	conference, proposal, paper, submission, session, presentation, submit, poster, deadline, topic	↑	
13	Travel information (0.019)	currency exchange, meeting, travel saving, transportation	san, travel, city, food, train, water, station, street, building, bus	-	
42	Grove music online (0.019)	grove, grove online, grove dictionary, new grove 2(ng2)	grove, online, article, dictionary, reference, print, oxford, grove_online, edition, resource	-	
15	Hymn (0.019)	Hymnals, hymn tune, chant, mass, choral music	church, organ, hymn, choral, saint, psalm, sing, choir, antoinette, mass	-	
49	Job posting (0.019)	Job posting	job, position, service, placement, librarian, apply, mla, placement_service, mla_placement, hire	↑	
24	CD/DVD sales (0.018)	CD HotList, Music Media Monthly, MLA Discount	order, sale, label, cd, offer, release, special, set, time, mla	↑	
27	Call for papers proposals, awards, etc. (0.014)	Call for Papers, Call for Submissions, Call for Proposals, Call for Applications, Call for poster sessions, call for seminar topics	letter, support, grant, application, travel, award, year, annual, meeting, moug	↑	

Table 2. Identified topics by topic modeling

that MLA members tended to engage more on other issues than these traditional library functions in recent years. This shift of attention may be attributed to advancement of technologies in recent years, including those in MIR. For instance, the downward trend of topic #12 (Requesting songs) might be related to the fact that music librarians and their clients have been equipped with alterative means to discover and access songs and other musical materials, such as search engines and online music repositories. Another trendline that may also reflect technology advancement is topic # 2 (music storage). The sharp decrease of this topic corresponds to the decline of CDs, LPs, tapes as formats of physical music materials. In this regard, technologies and resources facilitating music information retrieval and access are of great demand, particularly when the ways and channels people look for music information are changing so rapidly.

On the other hand, topics with upward trends include #20, which consists of messages about MLA board and members, demonstrating a vibrant and active professional association in the field of music librarianship. The second topic with a growing popularity is #31 (Metadata), which corroborates with recent developments in the metadata field such as RDA (Resource Description and Access, a standard for cataloging released in 2010). The other topics with upward trends all fall into the Advertisement category: #7 and #27 (Call for papers, proposals, awards), #49 (Job postings) and #24 (CD/DVD sales). This again reflects the flourishing development of the field and the community. In fact, the MLA and the U.S. branch of the International Association of Music Libraries (IAML-US) were merged in 2011, which has substantially boosted the status of MLA in the profession.[6]

6.3 Topics of Emails with/without Replies

We compared the email messages that stimulated discussion among participants of the MLA-L to those that did not. Figure 3 shows topics and their normalized weights in emails with and without replies respectively. Among the 27 topics identified by topic modeling, three pairs were merged for this analysis as the semantics of each pair are almost identical: #39 and #49 (Job postings), # 23 and #44 (Journals and periodicals) and #7 and #27 (Call for papers/proposals/awards). As shown in Figure 3, there is an overlap among highly-ranked topics between the two lists, such as "Cataloging", "Patron's request & questions", and "Classical music". This is not surprising as these are the most popular topics in the entire dataset (Table 2).

It is more interesting to see that there are topics ranked high in emails with replies but low in those without, such as "Subject heading, LC, genre, code" (#0), "Requesting songs" (#12), "Audio recording" (#47), and "Copyright" (#29)—indicating that emails in these topics often started discussions among subscribers. In particular, emails in "Requesting songs" (#12) and "Audio recording" (#47) are likely to contain music information needs and queries for music information. Similar to postings in Q&A websites, these email exchanges provide insights on 1) what kind of music information was needed by music librarians who in turn were trying to meet the needs of their patrons (i.e., the end users); and 2) how well-trained music information professionals looked for music information. Some of the heated discussions in threads with a large number of replies are likely to include queries that were interesting yet hard to find information for. These are excellent resources to discover not only new use cases but also search strategies for novel MIR systems.

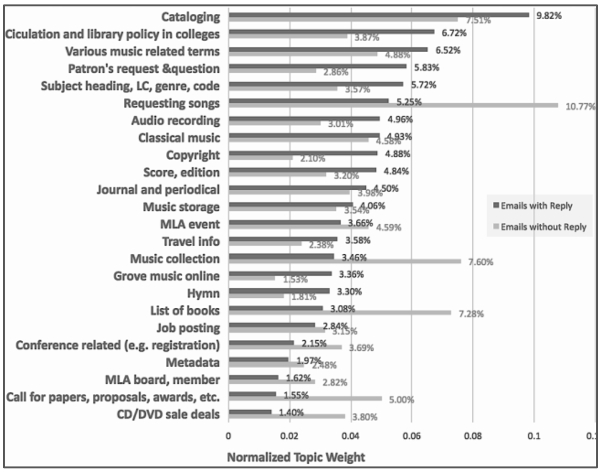


Figure 3. Topics in emails with and without replies.

Topics that are much more popular in emails without replies than in those with replies include "MLA event" (#4), "Call for papers/proposals/awards" (#7 and #27), "MLA board, members" (#20), and "CD/DVD sales" (#24). These topics are mostly of the nature of announcement and thus are unlikely to trigger discussions.

7. CONCLUSION

This study collected email messages posting in the Music Librarian Association mailing list (MLA-L) from 2000 to 2016, and analyzed the content through text mining. Main topics of discussions and their trends over the years are identified using Latent Dirichlet allocation (LDA). Twenty-seven meaningful topics were found and their semantics and trends were discussed in the context of MIR research. Topics in emails with and without replies were compared. As music librarians are gateways and bridges between music resources and users, the goals of music librarians and those of MIR researchers and practitioners are consistent: to help and facilitate users to access and make better use of music information. Therefore, the concerns and focuses reflected in the MLA-L are worthy of attention from the MIR community.

Future work will include detailed content analysis of the emails in such topics as "Requesting songs" and "Patron's Request & questions", to identify the needs of music information professionals and their users, and to learn about effective strategies of identifying and locating hard-to-find music information.

[6] https://www.musiclibraryassoc.org/?page=AboutMLA

8. ACKNOWLEDGEMENT

We thank the anonymous reviewers for their helpful comments. This study is partially supported by a seed research grant funded by the University of Hong Kong, and the Data-Driven Scholarship Autumn Institute jointly sponsored by the iSchool at the University of Illinois and the Faculty of Education at the University of Hong Kong.

9. REFERENCES

[1] D. Bainbridge et al., "How People Describe Their Music Information Needs: A Grounded Theory Analysis of Music Queries," in *Proc. of ISMIR*, 2003.

[2] A. Barua et al., "What Are Developers Talking About? An Analysis of Topics and Trends in Stack Overflow," *Empirical Software Eng.*, 19(3), pp. 619-654, 2014.

[3] D. M. Blei and J. D. Lafferty, "Dynamic Topic Models," in *Proc. of the 23rd Int. Conf. on Machine Learning*, Pittsburgh, PA, 2006, pp. 113-120.

[4] D. M. Blei et al., "Latent Dirichlet Allocation," *J. Mach. Learning Research*, vol. 3, pp. 993-1022, 2003.

[5] D. Campana, "Information Flow: Written Communication Among Music Librarians," *Notes*, ser. 2, 47(3), pp. 686-707, Mar. 1991.

[6] V. R. Carvalho and W. W. Cohen, "Learning to Extract Signature and Reply Lines from Email," in *Proc. Conf. Email and Anti-Spam*, 2004.

[7] K. Choi et al., "Topic Modeling Users' Interpretations of Songs to Inform Subject Access in Music Digital Libraries," in *Proc. of JCDL*, 2015, pp. 183-186.

[8] D. R. Cox and A. Stuart, "Some Quick Sign Tests for Trend in Location and Dispersion," *Biometrika*, 42(1/2), pp. 80-95, 1955.

[9] J. S. Downie and S. J. Cunningham, "Toward a Theory of Music Information Retrieval Queries: System Design Implications," in *Proc. of ISMIR*, 2002.

[10] M. M. Edwards, "A Content Analysis of the PUBYAC Discussion List," M.S. thesis, UNC, Chapel Hill, 1999.

[11] R. Griscom, Richard, "MLA-L at Twenty," *Notes*, vol. 65, no. 3, pp. 433-463, 2009.

[12] D. Hall et al., "Studying the History of Ideas Using Topic Models," in *Proc. Conf. Empirical Methods in Natural Language Processing*, 2008, pp. 363-371.

[13] X. Hu et al., "WiMIR: An Informetric Study on Women Authors In ISMIR," in *Proc. of ISMIR*, 2016, pp. 765-771.

[14] X. Hu et al., "A Framework for Evaluating Multimodal Music Mood Classification," *JASIST*, 68(2), pp. 273-285, 2017.

[15] X. Hu and J. S. Downie, "Improving Mood Classification in Music Digital Libraries by Combining Lyrics and Audio," in *Proc. of JCDL*, 2010, pp. 159-168.

[16] J. H. Lee, "Analysis of User Needs and Information Features in Natural Language Queries Seeking Music Information," *JASIST*, 61(5), pp. 1025-1045, 2010.

[17] J. H. Lee et al., "An Analysis of ISMIR Proceedings: Patterns of Authorship, Topic, and Citation," in *Proc. of ISMIR*, 2009, pp. 57-62.

[18] T. Martino. (2009). *Trend Analysis with the Cox-Stuart Test in R*. http://statistic-on-air.blogspot.com /2009/08/trend-analysis-with-cox-stuart-test-in. html

[19] A. K. McCallum. (2002). *Mallet: A machine Learning for Language Toolkit* [Online]. Available: http://mallet.cs.umass.edu

[20] A. McCallum et al., "The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email," in *Workshop on Link Analysis, Counterterrorism and Security*, 2005, pp. 33.

[21] G. Mishne and N. Glance, "Leave a Reply: An Analysis of Weblog Comments," in *3rd Annual Workshop on the Weblogging Ecosystem*, 2006.

[22] R. Neumayer and A. Rauber, "Integration of text and audio features for genre classification in music information retrieval," in *Proc. of ECiR*, 2007, pp. 724-727.

[23] J. Riley and C. A. Mayer, "Ask a Librarian: The Role of Librarians in the Music Information Retrieval Community," in *Proc. of ISMIR*, 2006, pp. 13-18.

[24] M. Schedl et al., "Music Information Retrieval: Recent Developments and Applications," *Foundations and Trends® in Inform. Retrieval*, 8(2-3), pp. 127-261, 2014.

[25] C. Tan et al., "The Use of Bigrams to Enhance Text Categorization," *Inform. Process. & Manage.,* 38(4), pp. 529-546, 2002.

[26] L. Troutman, "MLA-L: A New mode of Communication," *Fontes Artis Musicae*, pp. 271-281, 1995.

[27] J. W. Uys et al., "Leveraging Unstructured Information Using Topic Modelling," in *Portland Int. Conf. Management of Engineering & Technology (PICMET)*, 2008, pp. 955-961.