4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15104

Automated Information Retrieval from the Bibliographic Metadata: A Way to Facilitate the Systematic Literature Review

Marie Vítová Dušková[1], Martin Víta[2]
[1]Department of Marketing, Faculty of Business Administration, Prague University of Economics and Business, Czech Republic, [2]Department of Mathematics, Faculty of Informatics and Statistics, Prague University of Economics and Business, Czech Republic.

Abstract

The aim of this paper is to demonstrate the possible enrichment of the traditional procedure of bibliographic literature review using Natural Language Processing (NLP) methods – automated information retrieval. Our task was to conduct a systematic review of academic literature focused on the classical music audience research in the context of arts management and arts marketing. As a core base, we used bibliographic metadata, extracted from the Scopus database. The limits of the most commonly used methods of bibliographic analysis of the literature, which are co-citation analysis and bibliographic coupling, are well known. Therefore, we also used one of the NLP methods for metadata analysis, which allows automated processing of large numbers of texts to overcome these known problems. Thanks to this, we managed to obtain a higher granularity of the researched topics, to reveal emerging topics and to identify gaps in research. To the best of our knowledge, such an approach to the systematic literature review in the field of social sciences has not yet been applied.

Keywords: *Bibliographic Analysis; Natural Language Processing; Bibliographic Metadata; Art Audience Research.*

1. Introduction

Literature reviews play a crucial role in academic research in gathering existing knowledge and examining the state of the art. Among the many types of reviews that exist (from critical to post-publication reviews), systematic reviews of the literature are the most informative and scientific, however, only if they are consistently implemented and well justified (Paul et al., 2021).

It is common for scholars from the field of marketing and management to justify the search for a research question only on the basis of a cursory and narrative review of the literature (Linnenluecke et al., 2020). Unlike narrative literature review methods, meta-analysis, which is used in systematic literature reviews, allows us to statistically integrate and synthesize previous marketing and management research to prevent inconsistencies in the selection of documents included in the review and to create accumulated knowledge in given area.

As part of our long-term research focused on the classical music audiences, we analyzed published works during the period when, on the one hand, academic research in the field of marketing and art management developed (Colbert et al., 2014; Rentschler et al., 2006; Walmsley, 2019) and at the same time there was also a development and changes in the behavior and preferences of the audience (Prieur et al., 2013) (i.e., the development of the researched topic). In view of these facts, the goal of the systematic literature review was to 1) explore the scope of classical music audience research in the context of marketing and art management over time, 2) identify the most influential articles that were (or still are) the starting point of the research, 3) reveal current trends and perspectives in music audience research, and 4) identify unresolved issues and research subareas.

To achieve these goals, we performed two automated analyzes – bibliographic analysis and document affinity analysis.

2. Methods and Data

A systematic literature search requires a replicable, scientific and transparent process of evaluating existing knowledge (published in the literature) in order to minimize biases resulting from the random inclusion or exclusion of specific studies in the literature search process (Linnenluecke et al., 2020).

2.1. Data Collection

The basic precondition for a systematic review is the creation of a comprehensive or at least representative data set, which includes data on available research (Tranfield et al., 2003). Relevant documents for our research were first searched in the Scopus database using specific keywords and search strings. This search has been carefully documented. The result of this

search is a core base containing 188 documents. Subsequently, this set of documents was enriched with other documents found using Google Scholar based on reference analysis. The resulting file contains 257 documents. The dataset contains citation information, bibliographical information, the text of the abstract, keywords and references.

2.2. Bibliographic Analysis

Co-citation analysis is a bibliometric technique proposed by Small (1973), which aims to map the structure of the research field by analyzing groups of documents that are commonly cited together. The main disadvantage of co-citation analysis is that it is seen as an approach to the "past" of the research field, as it is more likely to capture older contributions and well-established researchers, rather than the current state of research. The papers in each cluster tend to share some common themes and are considered the basic knowledge base of the research area: the key concepts and methods on which the researchers build.

Bibliographic coupling can be interpreted as the opposite process to co-citation: two publications are marked as bibliographically paired if there is a third publication that is cited by both publications. Bibliographic coupling assumes that when two articles show similar bibliographies, they are likely to represent the same or at least related research topics. Because the citing documents are more recent than the cited documents, this method is suitable for examining newer contributions.

Both bibliometric analyzes were performed using the bibliometrix package programmed in `R` (Aria et al., 2017).

2.2. NLP Application -- Document Affinity Analysis

After performing data preprocessing, the corpus is represented as a standard document-term matrix – i.e., as a table whose rows correspond to documents (in our case papers), the columns then correspond to words (terms).

If a word does not appear in the document, then the value at the intersection of the corresponding row and column is set to zero, otherwise, a positive real number is used -- it expresses both the frequency of occurrence of the word in the document (“more frequent is more important'”) and frequency across the corpus (“words that are in a large number of documents are not so important'”) – more precisely, tf-idf weighting is used. Therefore, on each line of the matrix we find a vector that represents the given document (the components of the vector correspond to the dictionary we have available, which originated from the whole corpus). This way of representing entities belongs to vector representations.

Having a vector representation available for each document, we are able to measure the mutual distance of these documents, i.e., their vector representations. For this purpose, we use a standard cosine similarity – this expresses the affinity between each pair of documents.

This process implicitly leads to an undirected graph with weighted edges (vertices correspond to individual papers, the width of the edge/line between them expresses the degree of their similarity). For illustration we present a graph showing the similarity of authors – we do not present a graph showing similarity of documents for spatial reasons (Fig.1).

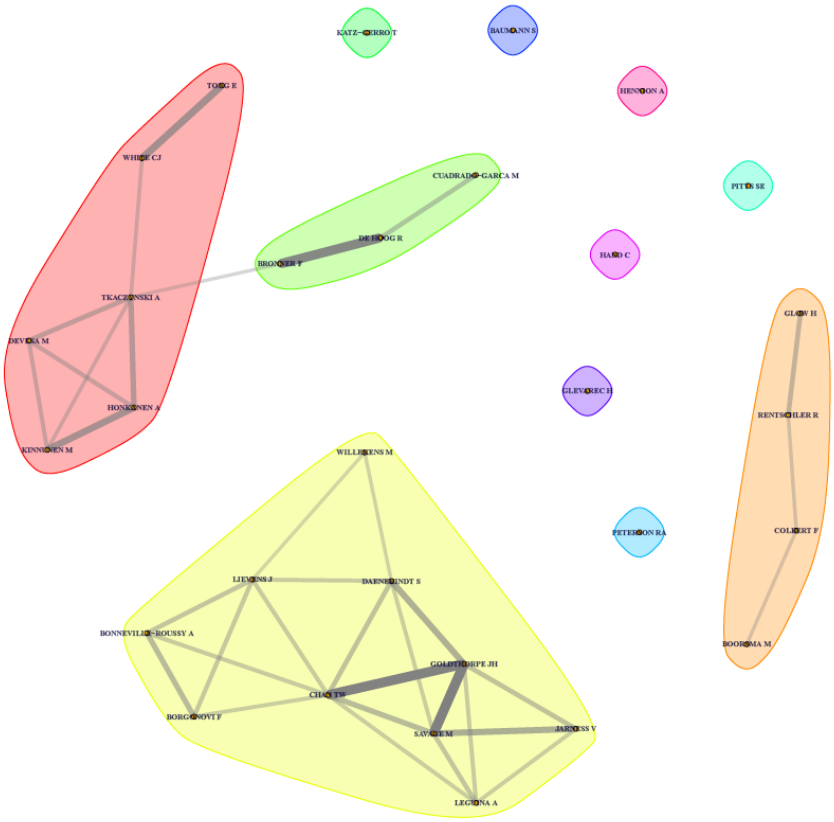

Figure 1 - Similarity of Authors

The WalkTrap algorithm implemented in `igraph` package (Sousa et al., 2014) was used to obtain clusters, i.e., communities of documents in our affinity/similarity graphs. As shown in Trigo et al. (2014), WalkTrap algorithm provides stable and useful results comparing to other approaches. The result can be represented in the form of a graph with highlighted sets of vertices. Formally, the output is a "key-value" table, in our case the title of the article (when analyzing the similarity of documents) or the name of the author (when analyzing the similarity of authors). Clusters can generally consist of a different number of elements (papers or authors), we will be particularly interested in those with the largest number of elements. Communities of authors are in fact induced by clustering (communities) of lists of

authors' papers. Hence the resulting author communities may contain authors with no common papers and no common citations etc.: they are grouped because they are related in the sense of working on the same or similar topic. This approach naturally extends approaches elaborated in Donthu et al. (2021).

3. Results

Each of the performed analyzes resulted in a number of clusters. In the case of co-citation analysis and bibliographic coupling, five clusters were created. In the case of document affinity/similarity analysis, more than twenty clusters were created. Upon closer analysis of the resulting clusters and after manual processing, their number was reduced to thirteen, because some clusters were very close thematically. Typically, four to six key references were published for each cluster; here, for space reasons, we only mention some of them (most cited in the bibliographic analysis, the most relevant in terms of the topic in the similarity analysis of documents).

3.1. Co-citation Analysis

The result of the co-citation analysis was five clusters, which were named according to the main topic of the documents included in the cluster. To save space, here are the cluster names and the key references for each cluster. More detailed papers' information, see the references in this paper.

- Cultural Capital
- Omnivorousness in Cultural Consumption
- Social Boundaries of Cultural Consumption
- Marketing Challenges in Audience Research
- Consumer Behavior

Among the key references for each cluster, we find very well-known texts and authors, e.g. Bourdieu (1984), Peterson (1992), Peterson et al. (1996), Lamont et al. (1992) or Van Eijck (2001).

3.2. Bibliographic Coupling

The bibliographic coupling resulted in five clusters. Their names correspond to the common topic of the documents included in the clusters.

- Cultural Stratification and Omnivorousness
- Musical Tastes and Musical Preferences
- Cultural Consumption Determinants, Arts Participation Boundaries
- Audience Research
- Festival Audiences and Event Marketing

Among the key references for these clusters, we alsofind very well-known texts and scholars, e.g. Peterson et al. (1996), Borgonovi (2004), Pitts (2005) or Bonneville-Roussy et al. (2013).

3.3. Document Affinity Analysis

The result of the document affinity analysis was thirteen clusters:

- Audience Development
- Audience Experience and Engagement
- Consumer Behavior
- Audience Segmentation
- Arts Marketing - Audience Research Theory
- Festival Audience
- Symphony Orchestra Audience
- Opera Audience
- Cultural Consumption and Social Stratification
- Arts Participation
- Tastes and Preferences
- Age and Musical Preferences
- Other Factors Influencing Arts Participation

At first glance, we see higher specialization of individual topics. In addition to older work, clusters include the latest work in the field, e.g. Soares-Quadros Junior et al. (2019), Daenekindt (2019), Vries et al. (2021), Kinnunen et al. (2019) and many others.

4. Discussion and Further Work

As we can see, the topics of the clusters from the individual analyzes overlap to some extent. Because the analyzes follow each other to a certain extent chronologically, we can see the development of the researched topics over time: in the citation analysis the topics are more general, in bibliographic coupling they are more specific in relation to our main research topic (i.e. classical music audience research) and document affinity analysis shows very specific research topics and directions. It is not without interest that, for example, the cluster "Other Factors Influencing..." includes only papers from 2014-2019, that means relatively new works.

The document affinity analysis divides all documents from the dataset into clusters, thanks to which we can then see the latest literature in the context of the older one. We have already mentioned that the main disadvantage of the bibliometric analyzes is that they provide a retrospective view of the researched field and draw attention to the most cited papers. To a certain extent, this can also lead to some distortions in future research – especially in research fields that are not widely exposed, there are certain communities of researchers in which there

is almost an obligation to cite some scholars, although their work may not be so crucial for specific research. Therefore, we see as a great advantage of document affinity analysis that it eliminates this influence – document affinity analysis only classifies documents into clusters based on their similarity and not importance within the scientific community.

Since affinity analysis seems to be a promising way to enrich "classical approaches"' to create systematic literature reviews, one of the directions of further research is to improve the computation of affinity by replacing tf-idf vector representations by state-of-the-art text embeddings arising from deep learning approaches (i.e., BERT).

Another direction is the development of a powerful visualization that incorporates both affinity results and co-citation/coupling results. Next step is then a development of a web based application, i.e., a web interface to our scripts in order to allow the user to create such reports and results interactively (without using our raw scripts exploited in this paper).

5. Conclusion

The aim of this paper is to demonstrate a possible approach to conducting a systematic literature search. Commonly used approaches – narrative literature processing on the one hand and bibliographic analysis on the other – may not always produce the desired results, especially if used in isolation. As we worked with a lot of data extracted from the web, we used tools that allowed us automated processing of such data. When processing a literature review, the researcher never avoids a certain amount of time-consuming manual work and careful study of found literature, yet our approach to the review makes it easier to discover research topics in the field (and to some extent quantify and visualize them in the context of other research topics), find research gaps or new research trends, thanks to the application of automated data processing.

References

Aria, M. & Cuccurullo, C (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of informetrics* 11.4, 959–975.

Bonneville-Roussy, A., Rentfrow, P.J., Xu M.K., & Potter J. (2013). Music through the ages: Trends in musical engagement and preferences from adolescence through middle adulthood. *Journal of personality and social psychology* 105.4, 703.

Borgonovi, F. (2004). Performing arts attendance: an economic approach. *Applied Economics* 36.17, 1871–1885.

Bourdieu, P (1984). Distinction: A Social Critique of the Judgement of Taste. *Cambridge, MA: Harvard UP*.

Colbert, F., & St-James Y. (2014). Research in arts marketing: Evolution and future directions. *Psychology & Marketing* 31.8, 566–575.

Daenekindt, S. (2019). Out of tune. How people understand social exclusion at concerts. *Poetics* 74, 101341.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133, 285-296.

Kinnunen, M., Luonila, M., & Honkanen, A. (2019). Segmentation of music festival attendees *Scandinavian Journal of Hospitality and Tourism* 19.3, 278–299.

Lamont, M. et al. (1992). Money, morals, and manners: The culture of the French and the American upper-middle class. *University of Chicago Press.*

Linnenluecke, M. K., Marrone, M., & Singh, A. K. (2020). Conducting systematic literature reviews and bibliometric analyses. *Australian Journal of Management* 45.2, 175–194.

Paul, J., Lim, W. M., O'Cass, A., Wei Hao, A. & Bresciani, S. (2021). Scientific procedures and rationales for systematic literature reviews (SPAR-4-SLR). *International Journal of Consumer Studies.*

Peterson, R. A. (1992). Understanding audience segmentation: From elite and mass to omnivore and univore. *Poetics* 21.4, 243–258.

Peterson, R. A. & Kern, R. M. (1996). Changing highbrow taste: From snob to omnivore. *American sociological review*, 900–907.

Pitts, S. E. (2005). What makes an audience? Investigating the roles and experiences of listeners at a chamber music festival. *Music and letters* 86.2, 257–269.

Prieur, A. & Savage, M. (2013). Emerging forms of cultural capital. *European Societies* 15.2, 246–267.

Rentschler, R., Radbourne, J., Carr, R., & Rickard, J. (2006). Relationship marketing, audience retention and performing arts organisation viability. *International journal of nonprofit and voluntary sector marketing* 7.2, 118–130.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science* 24.4, 265–269.

Soares-Quadros J., Fortunato, J., Lorenzo, O., Herrera, L., & Santos, N. S. A. (2019). Gender and religion as factors of individual differences in musical preference. *Musicae Scientiae* 23.4, 525–539.

Sousa, F. B., de, & Zhao, L. (2014). Evaluating and comparing the igraph community detection algorithms. *2014 Brazilian Conference on Intelligent Systems. IEEE*, 408–413.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British journal of management* 14.3, 207–222.

Trigo, L., & Brazdil, P. (2014). Affinity analysis between researchers using text mining and differential analysis of graphs. *ECML/PKDD 2014 PhD session Proceedings*, 169-176.

Van Eijck, K. (2001). Social differentiation in musical taste patterns. *Social forces* 79.3, 1163–1185.

Vries, R. de & Reeves, A. (2021). What does it mean to be a cultural omnivore? Conflicting visions of omnivorousness in empirical research. *Sociological Research Online*, 13607804211006109.

Walmsley, B. (2019). Understanding Audiences: A Critical Review of Audience Research. *Audience Engagement in the Performing Arts*, 25