Topic Models in Sport Management

Theodore Hayduk III, Texas A&M University
Andrew Pickett, University of South Dakota

Method - Multilevel analysis (Other)
20-minute oral presentation (including questions)
Abstract 2017-149
Friday, June 2, 2017
10:30 AM
Room: Torrey's Peak

Introduction
Topic modeling is a quantitative statistical method designed to generate structural meaning in large corpora of data (Blei, Ng, & Jordan, 2003; Reed, 2012). That is, in nearly any given field, there is an unmanageable amount of information for any single individual to consider and interpret. As examples, the entirety of the Encyclopedia Britannica or thirty years of published research in leading sport management journals could be a corpus of data for which greater organizational and thematic understanding may yield interesting insights. Topic modeling seeks to generate a thematic structure about the corpus of data in question, essentially mapping topics discussed in the selected documents. To date, only one article that we know of has used topic modeling in sport management (Shilbury, 2012). However, topic modeling offers a number of advantages in organizing and understanding data when compared to traditional content analysis or systematic review methods. Therefore, the current work argues for increased use of this method in sport management. Using a heuristic example, we present the topic modeling process, mapping the history sport management literature through an analysis of the conversations occurring in its top academic journals.

Conceptual Framework
Topic models are a family of techniques that help researchers sort through large collections of data using statistical analysis of unit based patterns. However, topic modeling is still a relatively new statistical technique, as it is reliant on modern computing power to simultaneously analyze such large amounts of data, given that a corpus may contain tens of millions of words (Blei, Ng, Jordan, 2003). At its core, topic modeling analyzes the co-occurrences of words in a corpus to assign themes (topics) to the data (Wang & Grimson, 2008). Conceptually, the generative probabilistic approach to developing latent structures about a set of words is similar to the practice of assigning latent structures to measured variable items in exploratory factor analysis. That is, the primary output of a single topic model would contain several lists of words which often co-occur. The program then generates a topic for each group of co-occurring words.

The underlying logic of topic modeling assumes that words are purposefully chosen by authors from a theoretical “bag of words.” For example, an individual writing about baseball is likely to draw from a collection including, “inning” or “pitch,” and would be unlikely to draw the word, “swimming.” Using a Bayesian approach, topic models “read” large amounts of text to derive these theoretically related bags of words, organizing terms based on frequent co-occurrence. However, this further occurs at multiple hierarchical levels, thereby adding more nuanced organization of terms. That is, simple word counts may group together articles about both baseball and ticket sales, both of which commonly use the term, “pitch,” but topic modeling would separate these as thematically different based on other words likely to occur with “pitch” in the respective documents. In an academic setting, these hierarchical levels may include an individual article, an edition of a specific journal, an entire journal itself, and the complete set of journals in a field. (Blei & Lafferty, 2007). Topic modeling is not limited to textual data. Because computers assign no inherent meaning to words aside from being a uniquely ordered combination of letters, the same logic could be extended to combinations of pixels and the underlying structures within a set of photographs could similarly be derived.

To this point we have described static topic models, which only evaluate data at a single point in time. That is, an entire corpus of data is treated as though it exists and was written contemporaneously. Of course, this is not the case. Dynamic topic modeling, therefore, subdivides a corpus into constituent time segments. This allows for the tracking of topics over time, noting the relative rise and fall of topic popularity in the corpus. Thus, a dynamic topic modeling approach could examine the relative inclusiveness of new topics by journals over time, as well as concentrations within the field in a given time period. Further, topic models divided by journal would allow for more accurate
definition of the breadth of topics discussed within a given journal and, accordingly, those that are not discussed in that publication. In a similar vein, dynamic topic models also allow for the examination of methodologies over time by topic (or by journal, etc). It may be interesting to note how quickly (or slowly) topics within a stream of literature evolve from qualitative to quantitative methodologies.

In a field as new as sport management, many of these findings may seem simple or intuitive to tenured authors who have significantly contributed to the field. However, topic modeling offers two primary benefits to more traditional content analysis and review methods. First, the use of computers to calculate topics saves authors’ time from manually reviewing the entirety of the corpus, thus extending the limit of feasibility. Second, and more importantly, topic modeling offers an impartial approach to assigning topic priority, thereby removing knowledge bias of a particular researcher’s findings. That is, it eliminates the risk that a tenured researcher who undertakes a content analysis of the entire field would over-represent her or his own subfield.

Method
This presentation will explain the process of dynamic topic modeling through a heuristic example, examining a corpus of data comprised of every article published in each of four journals, from their respective beginnings through 2015. The journals included are the Journal of Sport Management, Sport Management Review, Sport Marketing Quarterly, and European Sport Management Quarterly. Using Latent Dirichlet Allocation (LDA), we will first derive the common topics found in the field of sport management. Then, expanding these findings, using a dynamic topic model approach, we will examine these topics’ historical popularity in publication.

Expected Findings and Discussion
Given the relatively short history of the field, we expect that publication topics found in the first step will be largely intuitive. For example, we expect that general topics, such as marketing, strategy, diversity and inclusion, and organizational behavior will be present. Smaller subtopics, such as corporate social responsibility or LGBT inclusion may similarly present as salient clusters in the corpus. We expect the dynamic model to offer greater potential for thought-provoking insights. For example, we expect to see evolutionary effects within literature streams over time. For example, early discussions of diversity in sport may have related to race or gender, with LGBT issues becoming common only later in the corpus. Finally, by examining long-term and recent trends in the field, it may be possible to predict future salient topic areas. Using this context, we will both introduce and argue for increased use of topic modeling in sport management. We argue that it is time to begin the process of piecing together – in a consistent and validated way – the topics that have made up our past, comprise our present state, and will manifest in our future.