ADVANCES IN KNOWLEDGE ORGANIZATION
Proceedings of International ISKO Conference - ISSN: 0938-5495 - 2026 - São Paulo

# Dialogues between collaborative terms: the potential use of computational grounded theory (CGT) in the process of movie genre categorization

Rafael Rodrigo do Carmo Batista

Federal University of Santa Catarina, Brazil

Thiago Henrique Bragato Barros

Federal University of Santa Catarina, Brazil

## Abstract

The categorization of genre movies presents significant challenges for Knowledge Organization (KO) due to thematic hybridization and the opacity of commercial recommendation algorithms. This paper proposes a methodological framework for categorizing films by leveraging representative data from collaborative web databases, such as the Internet Movie Database (IMDb) and The Movie Database (TMDB). The study advocates for a shift from traditional classification to a more flexible categorization approach, utilizing user-generated content (tags, reviews, and overviews) as a source for "curated folksonomies." Methodologically, the research employs Computational Grounded Theory (CGT)—incorporating pattern detection, refinement, and confirmation—to process large volumes of audiovisual metadata. Technical procedures include using Python (Pandas/NLTK) for data extraction, TF-IDF and weighted metrics for indexing, and Sentence BERT (SBERT) with HDBSCAN for semantic clustering. Preliminary results, based on a corpus of 2,628 action movies, identify thematic clusters categorized as narrative, descriptive, or critical. The study concludes that integrating computational techniques with human-centered KO principles provides a robust alternative for identifying emerging themes and improving film representation in information systems.

## 1.0 Introduction

The categorization of genre movies is a highly complex issue for knowledge organization (Higgins 2022; La Barre and Cordeiro 2016). Based on the literature, the categorization of movie-related subjects attracts little attention precisely because of this complexity. Before reflecting on this, it is incumbent upon us, as KO theorists specializing in this type of document, to return to the point of origin and understand the transformations related to the distribution of these movies in a hyperconnected world.

We understand genre movies as a category of digital objects in the universe of audiovisual documents (Edmondson 1998). As an element, and when distributed widely through a commercial service, a movie becomes part of a collection whose thematic individuality is overwhelmed by a recommendation algorithm that capitalizes on the user's attention. Thus, these movies are silently associated with each other according to individual experience.

This entire process provides data that is constantly reused. In video-on-demand services, each piece of data is crucial for feeding back into this algorithm, making their platforms more robust for recommendations. However, where does the categorization of these movies by subject fit into this context, if these systems are often trade secrets? If the algorithm replaces curation, avoiding traditional search, I believe that the role of categorizing fiction film subjects could be associated with data on movies that, regardless of these algorithms, form true information communities.

Therefore, this paper proposes an organization of knowledge that leverages representative data on movies available on the web, from collaborative databases such as the Internet Movie Database (IMDb; https://www.imdb.com/) and The Movie Database (TMDB; https://www.themoviedb.org/). These networks allow users to express their impressions of movies at different levels through terms, reviews, and other forms.

The proposal here is to discuss the potential uses of this data, its advantages and limitations, by collecting the most frequent terms associated with two TMDB genre tags. Investigating this data can provide knowledge organization with a contextual perspective on the most relevant and emerging themes, as well as other thematic highlights over time.

To achieve this goal, we will first discuss the concept of categorization itself, connecting it to applications in film data research. Next, we will present CGT, indicating its nuances for qualitative data analysis. Finally, we will analyze alternative methods using data collected from action-genre movies in The Movie Database (TMDB).

## 2.0 Theoretical contributions on the use of collaborative data in movie categorization

Our understanding of categorization is like that of Schmidt and Wagner (2004), who view this process as the attribution of categories or concepts to phenomena through signs. The process of categorizing, in this way, consists of grouping the world into groups of entities with some common similarity (Jacob 2004). When considering an information system, categorization organizes its elements and establishes connections that make sense to the user. Jacob (2004) differentiates categorization from classification in three ways: its creative freedom, the presence of diffuse thematic boundaries, non-exclusive membership, and structure in varied clusters (which may or may not become hierarchies). This distinction is important because I consider categorization more appropriate for genre movies, precisely because of their potential for thematic hybridization. Higgins (2022) analyzes the difficulty of classifying movies in libraries using the Dewey Decimal Classification and the Library of Congress Classification and considers that classifying a film by subject (as a book) using these manuals does not facilitate the user's search process.

When considering online information systems, one's perspective changes. According to La Barre and Cordeiro (2016), Smeaton (2004), and Oberhelman (2014), methods for accessing movies on the web generally provide substitute representations (keywords, summaries, reviews) combined with metadata. The authors do not make any value judgments about these uses of representations, but their record indicates their validity. At the same time, they consider that access providers rarely incorporate the insights of film

theorists and other critical points of access, such as context, exhibition, and film production, even though research in Information Science and Library Science addresses the centrality of this access from the user's experience (Cordeiro 2010; Hertzum 2003).

One approach to representing film content focuses on the use of uncontrolled metadata, or tagging (La Barre and Cordeiro 2016). Neal (2012) indicates, however, that librarians and information scientists may not consider this type of free vocabulary the best option. La Barre and Cordeiro (2016) add that it is difficult to find cultural heritage through tags in web searches. This difficulty is related to the classic problems of folksonomy, such as the ambiguity of its terms, variation in granularity, synonymy, errors, homographs, and many personal or self-directed tags (Bullard 2018). One way to overcome these problems was proposed by Bullard (2018) through curated folksonomies, systems in which human judgment is preserved in the design of the folksonomy without sacrificing the scalability of computational processes. Unlike traditional folksonomies, curated folksonomies allow the production of controlled vocabularies, with the possibility of reviewing or refining the terms entered into a folksonomy system (Bullard 2018). The research proposal consisted precisely of using computational grounded theory to refine these terms.

## 3.0 Computational grounded theory as a tool for interpreting collaborative data on movies

Computational grounded theory (CGT) is a methodological framework created by Laura K. Nelson (2020) as a complement to the classic grounded theory of Glaser and Strauss (1999, 2005). According to the author, traditional GT faces the challenge of making the coding process for research data reproducible. To solve this problem, the author recommends applying computational techniques in three stages: pattern detection, pattern refinement, and pattern confirmation. The first stage involves the automatic extraction of text to categorize it, either by analyzing word frequency or through unsupervised text classification. According to Nelson herself, this first stage automates the initial coding process, allowing it to be reproduced even if the researcher uses specific criteria. In the second stage, the researcher returns to the data to confirm the plausibility of the previous stage, add their own interpretation, and, if necessary, modify the original patterns for better reading and understanding. The third stage verifies whether the patterns identified in the previous stages are generalizable to the entire corpus, ensuring the reliability of the process.

The method is interesting because it enables the use of big data tools in social science research. If we are analyzing a universe of X data on movies, the use of CGT has allowed us to work with a considerable volume of them. For example, using the pandas and NLTK libraries, it is possible to collect and structure data on 5,000 movies with genre tags from the Movie Database (TMDB) website. On TMDB, it is possible to extract overviews, reviews, and keywords using its main API. In my doctoral research, I have been working with this data from a set of movies.

## 4.0 An example using collaborative data from the action genre on TMDB

Within each of these classifications, clusters were characterized as weak or strong, depending on intracultural coherence—that is, whether the terms in a cluster appear in the corpus in a manner close to their original meaning, with low heterogeneity or reduced polysemy.

This classification was carried out through direct observation by the researcher, supported by its resemblance to the Word Intrusion logic tested by Chang et al. (2009), in which the quality of a topic is judged by the semantic homogeneity of its terms, i.e., the ease with which a human observer—in the context of this research, the researcher themselves—would identify intrusive terms.

Clusters with uniform terms in relation to their theme were classified as strong; those that presented artifacts, non-existent terms, semantic incoherence, or other problems were classified as weak. In terms of informational nature, a cluster can be narrative (when the prevalence is concentrated in overviews), descriptive (when the prevalence is concentrated in keywords), or critical (when the prevalence is concentrated in reviews). This predominance is determined by the proportion of term occurrences in each field of analysis and is also a product of the researcher's analysis of the obtained datasets. As an illustration, consider a cluster with three terms—A, B, and C—present, respectively, in 100 movies (70% in overviews, 30% in reviews), 50 movies (80% in overviews, 20% in reviews), and 200 movies (60% in overviews, 40% in reviews). The resulting weighted average points to a 65.7% predominance in overviews, classifying the cluster as narrative in nature.

## 5.0 Partial results

In compliance with the TMDB API parameters, data for 2,628 action movies were obtained during the pattern detection phase. These movies generated a total of 2,628 overviews, 12,713 keywords, and 509 reviews. At the end of pattern detection, 17 strong structuring clusters, 15 weak structuring clusters, 15 strong peripheral clusters, 5 weak peripheral clusters, 4 strong noise clusters, and 10 weak noise clusters were grouped.

First, as expected, the combination of semantic embeddings and density-based clustering produced some distortions. Some clusters obtained had a certain nature, but the terms that indicated it were “inflated” by polysemy or multicontextuality. The most problematic clusters were excluded from this stage. The final composition of these clusters revealed, in a preliminary interpretation, that the strong structuring clusters brought forward relevant themes in the films of the analyzed period. For example, 6 structuring clusters indicate a proximity of the collected action films to military or police themes. Likewise, there are sets of clusters that foreground relationships between characters as central elements of the films. Other sets deal with archetypes or essential characteristics of characters in this genre, as well as possible hybrid genres (the use of tags close to other genres marks action films that draw on the conventions of other genres).

The presentation of these initial results indicates that the division of categories in this proposal can provide valuable insights to professionals working in film categorization. The set of structuring clusters presents themes that support the existence of thematic

categories through narrative, descriptive, and critical clusters — the latter, however, presenting polysemic terms that require careful contextual interpretation.

The presence of clusters that mix genres further confirms that the boundaries for the categorization of these genres are, in fact, diffuse; the degree of importance generated through CGT can nonetheless hierarchize where these hybridizations occur, which could benefit manual categorization. Taken together, keywords, overviews, and reviews act as forces to be worked in conjunction with other categorization criteria rather than as substitutes for controlled categorization — a need for shared processes that points to the logic of the curated folksonomy proposed by Bullard (2018), regarding the preservation of the human factor without forgoing computational scalability.

AI Usage Statement: We declare that in this work we used Claude Code (Opus 5 model) to organize the code and utilize the TMDB API. The results were carefully reviewed by the authors prior to writing the paper.

References

Bullard, Julia. 2018. “Curated Folksonomies: Three Implementations of Structure through Human Judgment.” Knowledge Organization 45: 643–652. https://doi.org/10.5771/0943-7444-2018-8-643.

Chang, Jonathan, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. “Reading Tea Leaves: How Humans Interpret Topic Models.” In Advances in Neural Information Processing Systems 22, edited by Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, 288–296. Red Hook, NY: Curran Associates.

Cordeiro, Rosa Inês de Novais. 2010. “A recepção do espectador de filmes: parâmetros para a análise indexadora?” In Imagem, memória e informação, edited by Miriam Paula Manini, Otacílio Guedes Marques, and Nancy Campos Muniz, 81–106. Brasília: Ícone.

Edmondson, Ray. 1998. Uma filosofia de arquivos audiovisuais. Paris: UNESCO.

Glaser, Barney G., and Anselm L. Strauss. 1999. The Discovery of Grounded Theory: Strategies for Qualitative Research. New York: Aldine de Gruyter.

Glaser, Barney G., and Anselm L. Strauss. 2005. Awareness of Dying. New Brunswick, NJ: Aldine Transaction.

Hertzum, Morten. 2003. “Requests for Information from a Film Archive: A Case Study of Multimedia Retrieval.” Journal of Documentation 59: 168–186. https://doi.org/10.1108/00220410310463473.

Higgins, Colin. 2022. "'I Coulda Had Class' – The Difficulties of Classifying Film in Library of Congress Classification and Dewey Decimal Classification." Knowledge Organization 49: 79–86. https://doi.org/10.5771/0943-7444-2022-2-79.

Jacob, Elin K. 2004. "Classification and Categorization: A Difference that Makes a Difference." Library Trends 52: 515–540. https://hdl.handle.net/2142/1686.

La Barre, Kathryn, and Rosa Inês de Novais Cordeiro. 2016. "Film Retrieval on the Web: Sharing, Naming, Access and Discovery." In Managing Digital Cultural Objects: Analysis, Discovery and Retrieval, edited by Allen Foster and Pauline Rafferty, 199–218. London: Facet Publishing. https://doi.org/10.29085/9781783301539.010.

Neal, Diane Rasmussen. 2012. "Introduction to Indexing and Retrieval of Non-Text Information." In Indexing and Retrieval of Non-Text Information, edited by Diane Rasmussen Neal, 1–11. Berlin/Boston: De Gruyter Saur.

Nelson, Laura K. 2020. "Computational Grounded Theory: A Methodological Framework." Sociological Methods & Research 49: 3–42. https://doi.org/10.1177/0049124117729703.

Oberhelman, David. 2014. "Lights, Catalogues, Archives!: A Selected Annotated List of Web Resources for Film and Television Studies." Reference Reviews 28, no. 1: 2–5. https://doi.org/10.1108/RR-06-2013-0145.

Schmidt, Kjeld, and Ina Wagner. 2004. "Ordering Systems: Coordinative Practices and Artifacts in Architectural Design and Planning." Computer Supported Cooperative Work 13: 349–408. https://doi.org/10.1007/s10606-004-5059-3.

Smeaton, Alan F. 2004. "Indexing, Browsing and Searching of Digital Video." Annual Review of Information Science and Technology 38: 371–407. https://doi.org/10.1002/aris.1440380109.