Inferring Structure for Design: An Inductive Approach to Ontology Generation

Alfred Castillo[1], Arturo Castellanos[2], and Debra VanderMeer[3]

[1] CalPoly, San Luis Obispo, California, USA
[2] Baruch College (CUNY), New York, New York, USA
[3] Florida International University, Miami, Florida, USA

Abstract. We propose an approach for building hierarchical ontology structures based only on a set of category labels and knowledge of class membership. We illustrate our approach in the context of movies and provide a brief discussion on the theoretical and practical implications to design.

Keywords: Ontology, Basic Level Category, Recommendation Systems, Conceptual Modeling, Systems Analysis and Design

1 Introduction

The global market intelligence firm IDC estimates that more than 90% of enterprise data that are generated are unstructured (Gantz and Reinsel 2011). Discussion about structured or unstructured data have resurfaced recently as organizations try to create competitive advantage by analyzing and repurposing data (Schneider 2016). Some of the work in this area studies the tradeoff of using structured vs. unstructured formats. Structured data formats inherently provide the context for that data's use, however, unstructured formats can provide support not only to current use cases but also to future ones (e.g., in a web form, the user can be afforded flexibility on what data to enter rather than having to select from a pre-defined list of options in a drop-down menu) (Lukyanenko et al. 2014). Regardless of the choice, research has found that a mismatch between the IS class model (e.g., which would define the options in a drop-down menu) and the data the user wishes to enter can lead to low quality data or prevent users from contributing (Lukyanenko et al. 2014), and once the specification is made, the design choice can both serve to enable and constrain the IS (Hirschheim et al. 1995).

It seems logical that providing at least some degree of structure to data is beneficial as it provides context to the data points by design, and allows for differing various levels of expertise. Lukyanenko et al (2014) showed that expertise plays a major role in the "ideal" level of specificity in presenting users with options from a data quality perspective, in work using citizen-science data. Providing highly specific category options for bird watchers to log data, such as "Cygnus Atratus", as opposed to "Black Swan" or more broadly, "Swan", provides good data quality from the perspective of the experts, but the novices are left essentially guessing. Conversely, providing the

more friendly category of "Black Swan" may help a novice's selection be more accurate, but it is at the detriment of specificity (e.g. was it actually a Cygnus Atratus, or could it have been a Cygnus Melancorypus?). The level of specificity in designing a system that presents users with choices is therefore a very important consideration that requires care is taken to truly understand the population it is designed to serve.

Personalization of a system to serve multiple types of users is challenging. In order to present a user with a list of options, the system designer must decide on what areas within a taxonomy to present the user. A taxonomy is a hierarchical structure that organizes related information based on super/subset relationships from most generic (higher-order categories) to most specific. These categories effectively become the classes in the conceptual model. A subset of this taxonomy is usually presented to the user when organizing information, and system menus intended for the general population, such as e-commerce navigation, are designed for the most inclusivity by providing broader (higher-order) categories for their data. This achieves cognitive efficiency for the average user searching for a product, but it also can frustrate their expert user base with the lack of inferential utility (Parsons 1996; Roach 1978). For example, navigating Amazon for all competitor products to the niche Microsoft Surface computer. This product is often called a "2-in-1" device because it has characteristics of both a computer and a tablet. Finding it in the Amazon category hierarchy can be challenging, since it has no category of its own. Arguably, this is probably for the best as it presents a menu choice that works for the majority of their customers, however truly personalized systems should aim to reduce information overload and cognitive effort of their users by prioritizing information that is relevant for the user. In order to do this, the presentation of choice to the user needs to take into account what the user is looking for as well as the user's level of comfort within a given domain (or subdomain).

Larger-scale design challenges, where analysts face significant difficulty in deciding on how to organize and present information and choices to a variety of users and contexts, have also been described in the literature (Castellanos et al. 2016; Lukyanenko et al. 2017):

- System openness: How can we design for users with diverse backgrounds, education, and functional needs?
- Physical context: How can we flexibly design for a variety of physical contexts, including physical constraints?
- Cold-start problem: How can we generate recommendations when little is known about a user? Which categories are universal for the average user?

These design challenges are even more challenging when the design or decision-making space must be derived from unstructured text. Machine learning, natural language processing, and other data analysis methods have provided a means of identifying the entities described in unstructured text, both within documents (Feldman and Sanger 2007; Jordan and Mitchell 2015), and across a corpus of documents (Abbasi and Chen 2005; Abbasi and Chen 2008). However, there is little work addressing the challenge of adding structure over those entities. In this context, it is important to understand the relationships among data elements, i.e., the meta-structure. Ideally,

this would be a navigable structure, capable of representing hierarchical relationships and allowing a designer to easily locate the desired level of specificity, and prune away excess detail when needed.

In this work, we propose a first step in this direction. Specifically, we propose a method for inferring a hierarchical structure based on class membership (i.e., an object is member-of) and determining candidate classes with higher utility (i.e., categories that are highly relevant to the user), and provide a use case example based on data drawn from Netflix categories.

This paper is organized as follows. We describe the theoretical background and related work in Section 2. We then describe our method, as well as a similar method based on current literature as a comparative method, in Section 3.

2 Psychological Foundation and Literature Review

According to theories in psychology, categories support *cognitive economy* and *inferential utility* (Rosch and Lloyd 1978). These are two vital functions of organisms and one of the defining mechanisms of human cognition and behavior (Corter and Gluck 1992). These functions compete for the same limited cognitive resources of human memory, attention, and processing power. Cognitive economy is achieved by maximally abstracting individual differences among objects and then grouping them in categories of larger scope (e.g., *German-Shepherd* and *Labrador Retriever* belong in the category *dog*) (Fodor 1998). By storing only a few categories, humans can easily memorize identifying characteristics of the different objects in a class. This promotes communication efficiency and social interaction (Murphy 2004). In contrast, inferential utility refers to a category's usefulness in predicting the specific behavior or properties of objects within the category, based on similarities among category members. Here, it is not necessary to have observed every peculiarity of a given object, since knowing the object's category membership implies a set of behaviors and characteristics. Generally, emphasizing larger scope categories (i.e. higher cognitive economy) comes at the expense of capturing and communicating individual characteristics of objects that are unique, while emphasizing uniqueness (i.e. higher inferential utility) comes at the expense of having categories that are narrow (specific).

Rosch, Mervis, Gray, Johnson and Boyes-Braem (1976) argued that humans favor classes that are most capable of supporting these competing objectives of classification. Research found that objects are typically identified at a particular level of abstraction that is neither the most general nor the most specific possible (Jolicoeur et al. 1984) but an intermediate one coined ***basic level category*** (Rosch et al., 1976). These categories are generally at the middle level and are the most differentiated (Murphy and Brownell 1985). Objects at the subordinate (lower than basic) level need higher perceptual processing compared to that of basic level categorization (Jolicoeur et al., 1984) whereas middle-level categories are learned most quickly or can be named more quickly after they were learned (Corter & Gluck, 1992).

Building on the above theories, Corter and Gluck (1992) proposed a model of classification optimality and category utility (CU). This model is designed to directly

operationalize the tradeoff between cognitive economy and inferential utility in a way that adheres to widely held propositions about human cognition in psychology. They argue that the usefulness of a class is rooted in its ability to predict unobservable attributes (inferential utility) and optimize information processing and transfer (cognitive economy).

To summarize, classification theory in psychology amasses considerable evidence for the existence of classes that maximize agreement among people with different backgrounds, education, and functional needs. Coined "basic level categories", these categories have been shown to carry a multitude of benefits, resulting in a significant cognitive bias toward these categories.

Using psychology theory as the basis for better understanding the hierarchical relationship among categories, we extend Corter and Gluck's work to produce a data-driven ontology.

3 Method for Ontology generation and BLCs

We used probability-driven models to derive an ontology and identify categories at different levels of abstraction (superordinate, basic, and subordinate). To test these models, we used a secondary dataset based on Netflix's movie-category membership, where we can identify the mapping between category labels (as tagged by Netflix) and the movie objects within each category. Although we can see the list of movies and the categories these movies belong to, we cannot see the full structure of the hierarchy, (nor are we aware that one exists).

Given any category within a domain (e.g., genre), the existing members of a category (e.g., "Steamy Horror Movies") will share similar attributes and parents, unless the selected category is itself the root of the ontology, which has no parent, as it is already the most abstract categorization. Within this problem space, Corter and Gluck's work has been broadly applied to the problem of deriving ontologies. However, in our work, we found two major difficulties in applying these methods directly to our problem space.

First, when given a set of category labels with no discernible root, where we situate ourselves in an ontology is of prime importance. Assume an ontology for "Musical Instruments." What is the probability of occurrence for the "Brass Instruments" category? What happens to the base rate probabilities if we then change the ontology to be rooted at the broader "Objects" category, which has the next order categorization of "Tangible Objects" and "Intangible Objects"? The same problem exists whether we look at base rate probabilities of group memberships or of features of objects within an ontology. For this reason, the domain needs to be scoped for relevance.

Second, we cannot assume that categories live within non-overlapping branches of an ontology tree. In fact, many categories belong to multiple parent categories, so there are effectively multi-rooted trees naturally occurring in our categorizations. For example, consider Figure 1, which depicts the superordinate categories for the "TV Action & Adventure" category, which is a child of "Action," "Adventure," and "TV Shows."

Figure 1. Example Subset of Genre Ontology

Further, there is also the possibility that some categories serve to bridge these trees. For example, the members of the "Brass Instruments" category will include trumpets, trombones, and saxophones, among other instruments. The memberships of these instruments will shed light on what categorizations are relevant, even if tangentially so. For example, saxophones also belong to "Woodwind" instruments as they utilize a reed, which vibrates to produce the sound that is then amplified in the instrument. This raises an additional issue in determining which categories are relevant, given the objects' traits under consideration. This issue exists whether we look at object traits related to group memberships (i.e. belongs to "Woodwind" and "Brass") or feature possession (i.e. an object within "Brass" contains a "reed" for resonance) within an ontology.

Our broad intuition in developing this work is as follows: in order to determine the rate of occurrence for any categorization, especially when there are multiple roots, the objects' memberships in the broader context need to be examined by considering in each category as a target and finding categories relevant to it, essentially filtering everything else out. In this work, we have adapted the Category Utility metric to address our problem space, and we compare the resulting trees with a comparable trees produced using Information Gain.

3.1 Adopting Category Utility for Our Proposed Method

There has been much work in deriving relationships among concepts using the PageRank algorithm, such as AuthorRank, Y-factor, CiteRank, FutureRank, Eigenfactor, TextRank, and many others (see Yan et al. (2011)). Although this algorithm was originally considered for our purposes, we had to exclude this and other possibilities due to the need for directed relationships among elements. In order to iterate to an accurate PageRank for your nodes, it requires sharing a previously calculated PageRank equally among all outlinks (all nodes that the source node points to). Without knowing directionality, it becomes virtually impossible to converge to a meaningful Pag-

eRank value. Although PageRank has been adapted to use in the text-mining space (Mihalcea et al. 2004), directionality is easily derived as there is a logical ordering of concepts embedded within the English grammar. The arrangement of words in a sentence affords proximity and positioning metrics that can produce the necessary directionality among the constituent words or phrases. However, in our case we are associating entire categories without knowledge of the relationship's direction, and producing an undirected graph as a result.

In order to adequately attribute relative "importance" of the concepts related in this undirected graph, as PageRank is able to, we were forced to look elsewhere. Corter and Gluck (1992) posit that classes with the highest CU will also be most universal among all humans, since knowing and storing them provides the greatest value and therefore can be considered basic. The category utility function is calculated as follows:

$$max\ CU = f(c,F) = P(c)\sum_{k=1}^{m}[P(f_k|c)^2 - P(f_k)^2]$$

In this formula, a class c is defined by a set of objects o that belong in it. Each object is characterized by a finite feature (attribute) set $F = \{f_1, f_2, \dots, f_m\}$ which is captured in k. CU can be considered as the expected reduction of uncertainty due to communication of category information through some cue. The uncertainty is maximal when no category is present, and uncertainty is reduced the more "informative" the category becomes –but balanced by the frequency of the category. Although the CU provides us with a measure of usefulness for categories, there is also a probability distribution that can indicate how they should be connected. To illustrate, assume there are 100 "students". Of these, 80 are "undergraduate" students and 20 are "graduate" students. In our student example, then, we have the following:

- P(undergraduate | student): 0.80
- P(graduate | student): 0.20
- P(student | undergraduate): 1.0
- P(student | graduate): 1.0

Notice that the higher probabilities are for the children, given the parent. Children of a given category typically agree on the highest probability of group membership being attributed to the parents.

Using our secondary data, we adapted Corter and Gluck to assign CU weights across all categories. Although base-rate probabilities seem like straight-forward calculation, they are entirely dependent on scoping. For example, the likelihood of any swan being a black swan is much higher than the likelihood of any bird being a black swan. The denominator in the calculation is directly affected by scoping choice. In our case, we have multiple nested ontologies and the set of class memberships of the objects within any category should provide an adequate proxy for what subset of the larger ontology is relevant for calculating these probabilities. The tree generation process using adapted CU would be as follows:

1. Calculate the probability for each category in the domain as $P(c)$.

2. Create key-value pairs for each category, where the key is the category label, and the value represents a counter with an initial value of 0. For each movie category,

get the movies that belong to that category. Then, for each member movie, increment the count of movies that have that category label.

3. The key-value pairs are used to calculate the relative frequency for each of the k related categories, $P(c_k)$ for each category. This is simply a count of c_k's members divided by the total memberships for all categories summed.

4. $P(c_k)$ Situated at a category c (selected as our scoping root category), we then calculate the probabilities of all other k categories, given this category, $P(c_k \mid c).\, c_k$ This is computed for all possible combinations, and recorded in a k * k matrix where the row is c and the column is c_k.

5. The category utility score is calculated for the selected category as Corter and Gluck's algorithm above, substituting f_k with c_k where applicable.

6. To represent these structures visually, nodes are added into the network analysis tool Gephi (Bastian et al. 2009) along with their CU score as a "weight" attribute. Also, all edges are added between the nodes with $P(c_k \mid c)$ as a "weight" attribute to produce a tree.

7. This process is repeated for all categories to derive the relevant portions of the entire ontology.

3.2 Using Information Gain as a Comparison Method

Alternatives are difficult to choose due to the undirected nature of relationships in our use case. Following related work (Bouza et al. 2008; Zhang et al. 2002), we use information gain to find the most relevant "attributes" (i.e., categories) given a movie with a given category. Intuitively the information gain of higher order categories should approach 1. This is similar conceptually to inheritance, where a "student" object (subordinate class) is also a "person" object (superordinate class). However, this stochastically selected category could itself be a parent to other categorizations. For example, using the previous student categorization it is clear that an "undergraduate" (subordinate class) is also a "student" (superordinate class). Information gain's value for any subcategory, based on entropy, will depend on how many sibling categories there are at that level. Using our secondary data, the Tree generation process using Information Gain would be as follows:

1. For each movie category, get the movies that belong to that category.

2. For each category, create a key-value pair with the category label as the key and a counter as the value, assigned a default value of 0. Then, for each movie, increment the count value for every category label associated with the movie.

3. For each category, consider the target variable to be the category itself and consider all the categories retrieved from (2) as the features, once the target variable is removed.

4. For each category, calculate the information gain and create a decision tree and a variable importance plot. We used the Sci-kit learn library for Python (Pedregosa et al. 2011) to calculate the entropy and generate the decision trees.

3.3 Experimental Dataset

For this study, we selected Netflix data as a relevant secondary data to explore methods for deriving an ontology and identifying potential basic classes (i.e., classes with the highest category utility). The primary reasons for this are: (1) the complexity of the data, in that a category (genre) could belong to multiple other parent categories (i.e. "Slasher and Serial Killer Movies" could have the parent categories of "Thriller" and "Gore"), and (2) there may be multiple children within any given category, producing a non-binary tree. Although choosing this dataset provides the complexity in the structure we desire, it also presents a challenge in evaluation as, to the best of our knowledge, there is no pre-defined ontology that has been made publicly available for these data.

The opportunity to analyze multi-rooted trees in the Netflix data is not unique to the video streaming context. As technology improves, the boundaries between classes become blurry. Going back to the Microsoft Surface example from the introduction, although the Surface shares much in common with: (1) a "laptop" as it has substantial memory and other computing resources; (2) a "tablet" as it is designed with extreme portability as a key competency, and without the need for a keyboard (although offering a keyboard as peripheral in this market segment has become quite popular); and (3) a "notebook pc" as it has significant computing power, but is very lightweight. An iPhone has valid argumentation for, and against, each of these categories as well. For example, even the category of a "smart phone," which arguably has cellular service as a distinction, is no longer distinct to it as portable computing devices are now available with built-in 4G, LTE capabilities.

The dataset was collected from Netflix by programmatically iterating through all of Netflix's existing movie offerings organized by category (e.g., http://www.netflix.com/browse/genre/genreid, where genreID is the id code of the genre/category). For example, the URL for Action & Adventure (ID: 1365) is https://www.netflix.com/browse/genre/1365/?so=az, where the "az" denotes that the list will be sorted in alphabetical order. The anatomy of the Netflix Movie Recommendation System was retrieved from: https://www.whats-on-netflix.com/news/the-netflix-id-bible-every-category-on-netflix/. It is important to note that it seems that Netflix categorization is constantly in change, which is discussed in the limitations section of the manuscript.

The data collected was stored in a NoSQL document-oriented database. MongoDB was selected for its schema-on-write capabilities, which allows flexibility in storing data. This flexibility is important because a given video can exist in any number of categories. For example, Stranger Things (ID: 80057281) belongs to 75 of our identified categories, and House of Cards (ID: 70178217) belongs to 26. All of the genre memberships for the videos were stored as an array, resulting in a large videoID x genreID matrix of dummy coded values (1 if member, 0 otherwise). The final dataset consisted of 5,729 videos (e.g., movies, TV series, documentaries) and 24,610 distinct categories.

4. Preliminary Results

In this section we give a brief example that illustrates the derived hierarchical structure through our approach and the tree generated using the Information Gain method. In the interest of space, we focus on a single category, Spanish Dramas.

Information gain gives an idea of the overlap that exists between categories, generating an implicit ontology—similar, although computationally different, to CU in Corter and Gluck (1992). The algorithm was able to detect the importance of Dramas, Comedies, and other similarly related categories in the domain of Spanish Dramas. However, notice that it was not very useful in providing an accurate ontology that would reflect what one would expect to see for Spanish Dramas. For instance, from all categories displayed, possible candidates for roots would be Dramas and Spanish Movies, as they are logically the most generic. However, even if we ignore directionality, the closest categories to Dramas are Spanish Comedies, Cynical Spanish-Language Comedies, Spanish Movies, and Witty Spanish-Language Comedies (see Figure 2). These categories are logically not immediate children of Dramas.

In our domain, membership of a category to a super-ordinate category is not mutually exclusive to other super-ordinates at the same level (i.e., any given genre can belong to multiple immediate parents) as illustrated in the logical extension of these movies belonging under Comedies as well as Dramas (see variable importance Figure 3). Although information gain is useful for determining if a given category is a parent (low entropy) or a subordinate (higher entropy) categorization, it will not indicate to us at what level the perceived parent is within the ontology. Although it may detect the parent, it does not necessarily place them appropriately at the right level (see Figure 2 and 3).

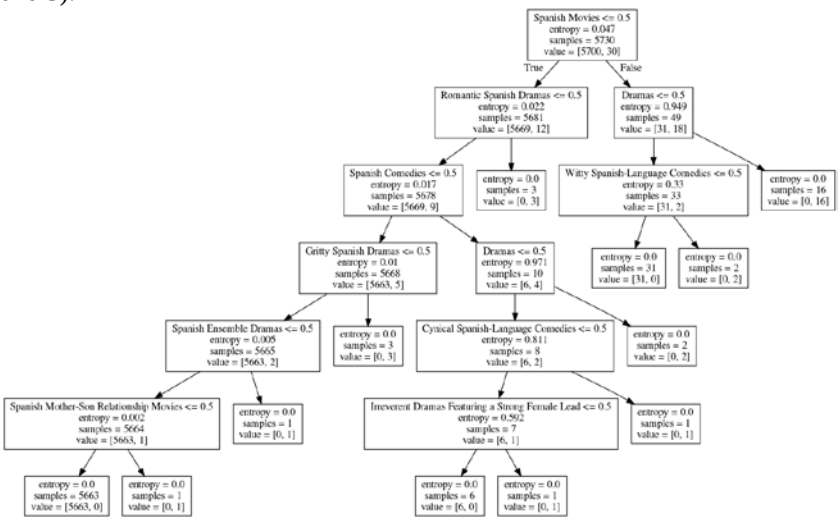

Fig 2. Decision Tree for the category "Spanish Dramas"

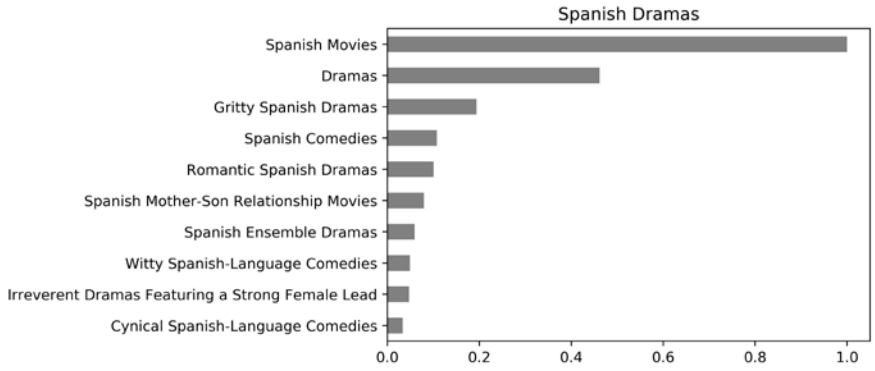

Fig 3. Variable importance for the category "Critically-acclaimed 20th Century Period Pieces"

Consider the same categories as rendered by our approach in Figures 4 and 5. Figure 4 is a zoomed-out view of the navigable structure produced by this method, while Figure 5 zooms in to demonstrate the similarities and differences, compared to the information-gain-based case. In this structure, the weights of edges are represented by color graduation from light green to dark green, and the weights of nodes are illustrated in the relative sizes of the label font and node circle. In this example, the closest and most meaningful categories detected were Dramas, followed by International Dramas, Thrillers, International Comedies, Comedies, and Critically-acclaimed Movies. This makes sense given the scope of the domain (i.e., Spanish Dramas). Aligned to the results from the Information Gain method, comedies are relevant to this category, although from a tangentially related tree (Comedies tree). Therefore, the closest and most relevant category from our generated ontology from the Comedies portion of the ontology is International Comedies, as these are Spanish films, which is followed by Comedies as important but slightly less relevant.

Notice also that our approach was also able to detect the tangentially related Thrillers category. Examples of movies in the Spanish Dramas category overlapping with Thrillers are the critically-acclaimed Toro, Black Snow, Smoke & Mirrors, Orbiter 9, and others. The Thrillers portion of the ontology that is relevant for this categorization was ignored in the Information Gain result. The same could be said for Critically-Acclaimed Movies (Toro as an example) and International Movies (they arguably all are tangentially related to this category as well). Our tree is arguably much more complete and representative of the target category's ontology than the Information Gain approach was able to provide.

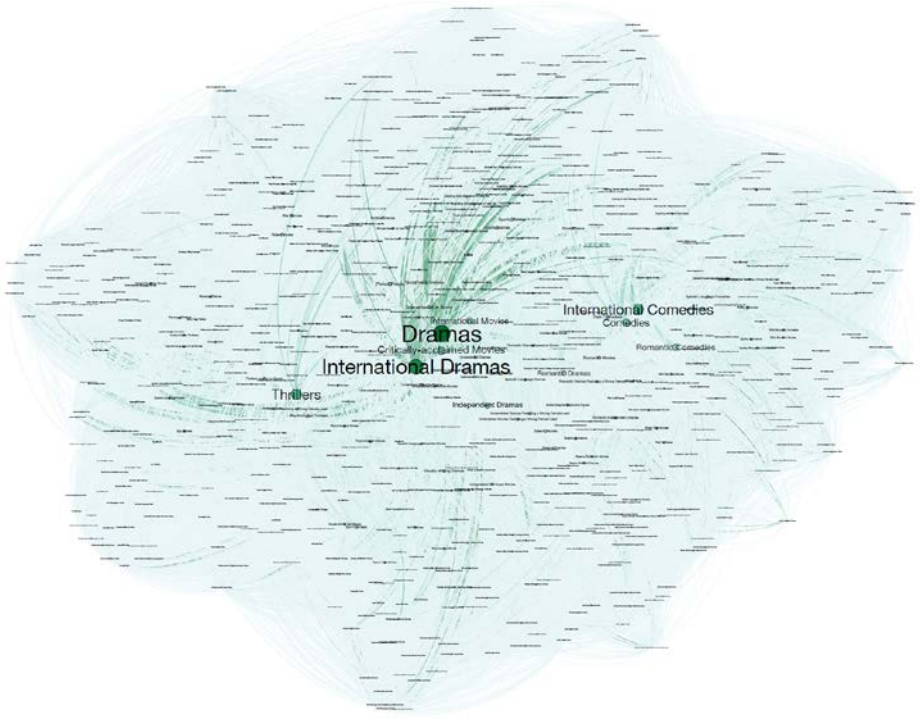

Figure 4. Scoped Ontology – Spanish Dramas

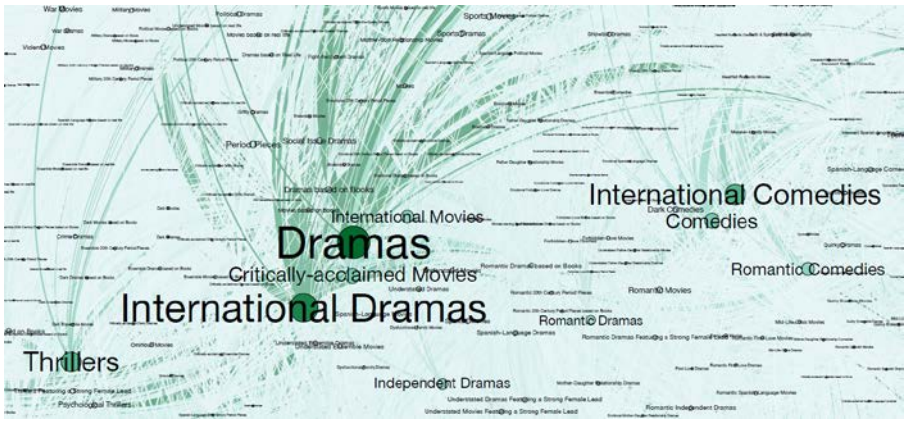

Figure 5. Zoomed in Scoped Ontology from Figure 4 – Spanish Dramas

Interestingly, the adapted Corter & Gluck method provides a serendipitous result (e.g., shows dramas and thrillers in the ontology) compared to the information gain method, which is less accurate as there may not be enough objects (e.g., movies) with certain characteristics within the sampled data. The algorithm uses membership (feature presence) to calculate its category utility—a proxy for relatedness in terms of

utility of a category within a structure. It is assumed that categories are created to the extent that they are useful (fulfill a functional role) and their ability to predict features of a member of this category. Basic level categories typically have higher category utility than both a superordinate and a subordinate category as there is an increase in predictive power from knowing that the object belongs to a category of which more inferences can be made (Tanaka and Taylor 1991). Corter and Gluck (1992) verified that category utility correctly predicts basic level categories.

4. Discussion

Classification theory in psychology provides evidence for the existence of classes that maximize agreement among people with diverse backgrounds, education, and functional needs. In this study, we used two probability-driven methods to derive an ontology from object membership and identify categories at different levels of abstraction (superordinate, basic, and subordinate). These categories vary as a result of the multi-rooted and non-binary nature of object categorization.

We explored ways that would help an analyst derive conceptual models and ultimately be able to present information at different levels of abstraction with the intention of providing users with information from the structure personalized to their needs, whether they need more detailed and specific, basic level information, or less detailed and less specific. Our method allows for hybrid categories which bridge ontologies together, as can occur in real data sets.

We believe our method contributes to the area of systems analysis and design (e.g., conceptual modeling) and has practical implications in that it would help analysts: (1) alleviate the diversity of the user base in open systems—by showing classes that are relevant to both experts and novice users, preventing users with low level of expertise to disengage from using the system (similar to the findings of Lukyanenko et al. (2014)); (2) deal with physical constraints (i.e., leverage basic level categories as candidate classes in a conceptual modes), and cold-start problem (i.e., use basic classes as anchors when we do not have enough information about the user).

We demonstrate our prototype method and derived a relevant ontology (from Netflix), as a proof of concept, anchored on a given category (“Spanish Dramas”), which is part of a broader unknown ontology.

The paper is not without limitations. Movie categorization may change over time (i.e., additional categories are added to movies). These variations could be a reflection of inconsistent or evolving tagging methods. Although we were able to show small sections (“pruned”) from the ontology, a method of piecing together all these ontologies should be explored. Although our focus is on the member-of relationship in order to generate a hierarchical relationship among data points, there are various meta-structures that could also be analyzed in future work. Future work will incorporate an evaluation method (that involves human judgment) to further support and validate the utility of the method beyond our proof-of-concept.

References

Abbasi, A., and Chen, H. 2005. "Applying Authorship Analysis to Extremist-Group Web Forum Messages," *IEEE Intelligent Systems* (20:5), pp. 67-75.

Abbasi, A., and Chen, H. 2008. "Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace," *ACM Transactions on Information Systems (TOIS)* (26:2), p. 7.

Bastian, M., Heymann, S., and Jacomy, M. 2009. "Gephi: An Open Source Software for Exploring and Manipulating Networks," *Third international AAAI conference on weblogs and social media*.

Bouza, A., Reif, G., Bernstein, A., and Gall, H. 2008. "Semtree: Ontology-Based Decision Tree Algorithm for Recommender Systems," *Proceedings of the 2007 International Conference on Posters and Demonstrations-Volume 401*: Citeseer, pp. 106-107.

Castellanos, A., Lukyanenko, R., Samuel, B. M., and Tremblay, M. C. 2016. "Conceptual Modeling in Open Information Environments," *AIS SIGSAND Symposium*, pp. 1-7.

Corter, J. E., and Gluck, M. A. 1992. "Explaining Basic Categories: Feature Predictability and Information," *Psychological Bulletin* (111:2), p. 291.

Feldman, R., and Sanger, J. 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge university press.

Fodor, J. A. 1998. *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press.

Gantz, J., and Reinsel, D. 2011. "Extracting Value from Chaos," *IDC iview* (1142), pp. 1-12.

Hirschheim, R., Klein, H. K., and Lyytinen, K. 1995. *Information Systems Development and Data Modeling: Conceptual and Philosophical Foundations*. Cambridge University Press.

Jolicoeur, P., Gluck, M. A., and Kosslyn, S. M. 1984. "Pictures and Names: Making the Connection," *Cognitive psychology* (16:2), pp. 243-275.

Jordan, M. I., and Mitchell, T. M. 2015. "Machine Learning: Trends, Perspectives, and Prospects," *Science* (349:6245), pp. 255-260.

Lukyanenko, R., Parsons, J., and Wiersma, Y. F. 2014. "The Iq of the Crowd: Understanding and Improving Information Quality in Structured User-Generated Content," *Information Systems Research* (25:4), pp. 669-689.

Lukyanenko, R., Wiersma, Y., Huber, B., Parsons, J., Wachinger, G., and Meldt, R. 2017. "Representing Crowd Knowledge: Guidelines for Conceptual Modeling of User-Generated Content," *Journal of the Association for Information Systems* (18:4), p. 297.

Mihalcea, R., Tarau, P., and Figa, E. 2004. "Pagerank on Semantic Networks, with Application to Word Sense Disambiguation," *Proceedings of the 20th international conference on Computational Linguistics*: Association for Computational Linguistics, p. 1126.

Murphy, G. 2004. *The Big Book of Concepts*. MIT press.

Murphy, G. L., and Brownell, H. H. 1985. "Category Differentiation in Object Recognition: Typicality Constraints on the Basic Category Advantage," *Journal of Experimental Psychology: Learning, Memory, and Cognition* (11:1), p. 70.

Parsons, J. 1996. "An Information Model Based on Classification Theory," *Management Science* (42:10), pp. 1437-1453.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. J. J. o. m. l. r. 2011. "Scikit-Learn: Machine Learning in Python," (12:Oct), pp. 2825-2830.

Roach, E. 1978. "Principles of Categorization," in *Cognition and Categorization*, E. Roach and B.B. Lloyd (eds.). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Rosch, E., and Lloyd, B. B. 1978. "Cognition and Categorization,").

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. 1976. "Basic Objects in Natural Categories," *Cognitive psychology* (8:3), pp. 382-439.

Schneider, C. 2016. "The Biggest Data Challenges That You Might Not Even Know You Have." IBM.

Tanaka, J. W., and Taylor, M. 1991. "Object Categories and Expertise: Is the Basic Level in the Eye of the Beholder?," *Cognitive psychology* (23:3), pp. 457-482.

Yan, E., Ding, Y., and Sugimoto, C. R. 2011. "P-Rank: An Indicator Measuring Prestige in Heterogeneous Scholarly Networks," *Journal of the American Society for Information Science and Technology* (62:3), pp. 467-477.

Zhang, J., Silvescu, A., and Honavar, V. 2002. "Ontology-Driven Induction of Decision Trees at Multiple Levels of Abstraction," *International Symposium on Abstraction, Reformulation, and Approximation*: Springer, pp. 316-323.