Plot Archeology: a vector-space model of narrative structure

Benjamin M. Schmidt
Department of History; NuLab for Texts, Maps, and Networks
Northeastern University
Boston, Massachusetts, USA
Email: b.schmidt@neu.edu

Abstract—A novel and important corpus of about 80,000 television and movie scripts from opensubtitles.com shows interesting large-scale patterns of narration in their vocabulary use. These patterns are interesting at the token level but not easily amenable for large scale data analysis. This paper describes a new method, “plot arcs,” for describing and comparing structural elements of structure, including plot, across large textual corpora by treating texts as paths through a multidimensional space derived from a topic model. Plot arcs offer a framework for describing the structure of text documents that is easily extensible to a variety of genres and can accommodate many different ideas of plot structure.

Index Terms—Plot, Topic Modeling, Document Structure, Television

I. DESCRIBING PLOTS

Scholars in the digital humanities are increasingly interested in methods for describing the ways that plots and other large-scale structural elements unfold. Although individual plots clearly form the heart of storytelling, this is not a topic for which traditional information-retrieval tools present obvious choices. The presence or absence of plot archetypes is primarily a matter of exploratory data analysis, but heuristics that are useful for exploring plot types are difficult to define and validate. Tools from machine learning and statistics offer some hope of opening new avenues for exploring plot structures at scale through the analysis of large corpora of narrative, but the relevant features and transformations that are constitutive of “plot” will necessarily always be vague.

A number of methods to explore regular aspects of plot have been suggested in the past few years. Matthew Jockers proposes using sentiment analysis to trace fluctuations in plot to operationalize a typology first proposed by Kurt Vonnegut. [1][2]. Andrew Piper uses structural similarities within individual novels (as opposed to externally trained supervised models, such as those used by Jockers) to identify occurrences of a single type of plot, the conversion narrative. [3]

While the locus classicus for the study of plot in the humanities has been the novel, questions of textual structure extend well beyond it. There are, in fact, strong reasons to avoid the novel in describing plot. Novels are both long and of irregular length, raising complicated issues of around appropriate methods for smoothing. [4] Other research has applied this observation in practice. Nick Beauchamp, for example, uses Principal Components Analysis to track the trajectory of individual speeches. Reiter et al. [5] apply a bevy of methods on folk tales.

This paper uses a novel corpus of approximately 80,000 transcripts of movies and television shows, described below. These provide a relatively and clean and consistent source base for studying plots. For scheduling reasons, most television shows are of a fixed length (in the United States, generally 30 minutes or an hour, including time for commercials) and can thus be directly compared to each other. Structural analysis of plots is intrinsic to the genre’s own developement: the industry has its own history of segmenting and studying stories by plot developments that dates to the the late 1930s. [6] Insofar as regularities of plot exist, they should be easier to find in a field where structures have been enforced by organizational constraints.

A. Why study plot?

Before embarking further, it is worth asking what uses humanists might have for quantitative methods of describing plot. One potential application is as an advanced form of information retrieval; there may be use cases for identifying plots that meet a particular trajectory in a broader context (for instance, comedies that appear particularly “Shakespearian” or reality shows that draw on the structure of the crime procedural). An additional result might be the creation of trend lines over time to map things like the regularization of three-act structures in films, or the emergence of genres based on structural features as opposed to topical ones.

Ultimately, as in many humanities, these methods may find use as elements of a critical appreciation. Structural considerations are particularly well developed in the fields of poetry and music, where they can act most usefully as a vocabulary to describe what is distinctive about particular works of art and to complicate traditional accounts of structure. [7]

II. DATA AND PRE-PROCESSING

This paper studies plot by using a corpus of approximately 80,000 files in the ‘srt’ format provided by the website opensubtitles.org. Each file, (hereafter called a ‘script’) transcribes the dialogue from a single movie or television episode for use as subtitles; each block of text in the srt files is identified by the
period of time it should appear on the screen. These are crowd-curated files designed to be easily synced with video files. Inclusion in the corpus requires an individual to have uploaded files to the web site; the selection biases that this creates are not fully understood. The fidelity is fairly good, compared to other large collections of texts (such as OCR scans of novels), but there are occasional irregularities including differing rates of description of onscreen activities, capitalization, and translation of words from foreign languages. Speakers are not differentiated. All scripts in this set are in English, although in some cases the on-screen spoken language is different.

Each script can be linked through identifiers to IMBD.org, one of the most extensive metadata repositories in existence. It contains far more detail, in particular, than the library catalogs used by most digital humanities research; this opens up many more opportunities for verification than is possible when working with novels. In particular, IMDB contains a wide variety of tags that make it possible to discriminate between comedy, drama, crime, and other genres. For the structural analysis in the second half of this paper, the scripts are further reduced down to only those that can be identified as television episodes by this cross-referencing. This smaller corpus constitutes about 27,000 scripts.

Most models of plot rely on comparing sections of a work. For commercial television shows like many of those explored here, the ideal formatting would involve commercial breaks, but those cannot be extracted from the metadata and are inapplicable to movies and non-commercial television. Instead, each script is divided into a number of chunks. Each script is initially divided into three parts; each chunk is recursively split in two until the chunk size is between two and four minutes. The purpose of this recursive chunking is to facilitate comparisons between shorter and longer television shows by comparing fractional lengths, while having a constant fixed length of individual “documents” for topic modeling. An obvious alternative, dividing each script into chunks of a fixed time and then determining which 6th each falls into, is inappropriate because many common television show lengths will divide unevenly. Including the subdivisions, this process yields a corpus of approximately 1.8 million documents of, on average, 364 words apiece, with a variety of metadata from IMDB. This data is then stored in a Bookworm instance for easy interactive data exploration. Several exploratory charts are available online at http://bookworm.benschmidt.org.

This chunking on its own reveals a striking degree of difference across time for many words or phrases. Phrases like “forgive” and “love you” spike up noticeably at the end of scripts. Place locations like “school” and “office” become less common through the course of scripts. Somewhat more surprisingly, a number of words, such as “journey,” are most common at the beginning and end but less in the middle. Each of these patterns shows various regularities in the topics discussed in storytelling across tens of thousands of episodes; various tropes of plot are evident such as the trend towards resolution and reconciliation at the end of an episode of television or the final scene of a movie.

### A. Topic Modeling as dimensionality reduction

These individual lexical patterns are difficult to describe at scale because there are tens of thousands of individually meaningful words. Even for computational analysis, this produces not-insignificant amounts of data. The object of study can be thought of as a three-dimensional variant on the traditional term-document matrix. The first two dimensions remain terms and documents; the third dimension corresponds to position in the text. Exploring the general characteristics of this space is somewhat intimidating because of the size of the matrix. As 8 byte integers, the full movie array would require approximately 2 terabytes; the sparse version used in the bookworm takes 2.7 gigabytes.

There may be good reason to study the matrix as a whole, but dimensionality reduction techniques can help to make the object of study more computationally tractable for exploratory data analysis. Latent Dirichlet Allocation can be used as a form of dimensionality reduction that dramatically reduces the complexity of the problem. [8]

Using the software package Mallet [9], each token in the source document is assigned probabilistically to a topic. We can then treat the vocabulary space as one not of words but of topics. This serves two purposes at once; it makes the overall memory requirements several orders of magnitude smaller, and it makes the space more cognizable to analysts seeking patterns of plot regularities in a large corpus. After some experimentation with various sizes and stopword lists, a 127-topic model was trained using an extended version of the stopwords published online as a supplement to Jockers’ Macroanalysis. [10] For this training, all scripts, including movies, were used.

<table>
<thead>
<tr>
<th>topic</th>
<th>topic_label</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Wait wait minute Let’s Look Hurry let’s</td>
</tr>
<tr>
<td>10</td>
<td>clean smell water wash use bath bathroom</td>
</tr>
<tr>
<td>11</td>
<td>film movie show TV movies scene play</td>
</tr>
<tr>
<td>31</td>
<td>hair funny look joke laugh big teeth</td>
</tr>
<tr>
<td>34</td>
<td>sir course Thank dear London quite Ah</td>
</tr>
<tr>
<td>60</td>
<td>talk talking Look crazy understand Listen problem</td>
</tr>
<tr>
<td>61</td>
<td>animals animal bear food wild hunting lion</td>
</tr>
<tr>
<td>74</td>
<td>hear voice heard sound radio noise listen</td>
</tr>
<tr>
<td>78</td>
<td>game play ball team playing win football</td>
</tr>
<tr>
<td>80</td>
<td>girl girls boy look name beautiful pretty</td>
</tr>
<tr>
<td>85</td>
<td>drink wine beer drunk bottle drinking glass</td>
</tr>
<tr>
<td>90</td>
<td>years world land water sea ago life</td>
</tr>
<tr>
<td>110</td>
<td>Madame de Monsieur French dear course evening</td>
</tr>
<tr>
<td>113</td>
<td>God cool Whoa Look look dude Wow</td>
</tr>
<tr>
<td>119</td>
<td>married wife wedding husband love marriage woman</td>
</tr>
<tr>
<td>121</td>
<td>Agent agent security FBI team CIA agents</td>
</tr>
</tbody>
</table>

To illustrate the sort of topics pulled on this set, a randomly selected sample of topics is shown. Each is identified by the top 7 words, in order, in each topic. In general, they appear to show general topical coherence related to topics like “drinking” (#85), “exotic animals” (#61), or “marriage” (#81). Others are characteristic of certain types of dialogue, such as “Wait wait minute Let’s” and “God cool Whoa Look dude”. The use of 2 to 4 minute chunks of dialogue makes the topics more concrete than those sometimes pulled from literary sources. Although the topical coherence is reassuring...
and makes interpretation easier, the use of LDA as dimensionality reduction means that strict semantic coherence is not necessary. An alternative form of reducing the size of the term-document matrix, such as principal components analysis, might work as well.

B. Topics show trends in screen time

Many of these topics show marked trends in their usage in “screen time” (that is, in time defined by twelfth of the TV show or movie). These trends can be linear (stronger at the end or beginning of scripts) or arced (strongest in the middle and weakest at the beginning and end.) Simple linear correlation on each of the 127 topics can extract the topics that show the strongest linear (Fig. 1) and arcing (Fig. 2) trends. These are the ones that show the most promise as differentiators of plots.

In aggregate, these make it possible to make a number of generalizations about the structure of television shows. Families and rules for competition appear mostly at the beginning and end; drinking, sexuality, friends, and animals all appear more often towards the middle. Death becomes a more prominent theme late in most shows, as do court trials and the language of apology; schoolrooms, workplaces, and homes are all more frequently mentioned at the beginning.

III. CHARACTERIZING PLOTS AS A TRENDS IN TOPIC SPACE

The foregoing examples show the trajectories of individual words or topics across all scripts in the corpus. Although this provides an interesting way to explore certain general trends of fictional writing, particularly across time, there is reason to want the primary object of study to be plots, rather than their individual components.

The simplest way to do this is is to characterize each television show according to its topical breakdown. For example, the plot of the television show *Law and Order* can be expressed as a function of the most frequently represented topics within it by a streamgraph that shows the pattern rigorously adhered to by the shows’ scripts; a crime is investigated by the police in the first half-hour, and then taken to trial by criminal attorneys in the second half. (Fig. 3)

Direct comparisons between shows are made easier by thinking of individual scripts as traversing a space defined by all 127 topics. The concept of such a vector space may be unfamiliar to many scholars of literature, so an extended two-dimensional demonstration may be helpful. Consider the topics found above to have strong linear trends (Fig. 2). Three of them decrease in frequency over the course of the aggregate television scripts and semantically seem to bear strong associations to particular places: “house room place
live living apartment move,” “school class teacher college high kids year,” and “work job working office boss day company.” Three others descend and seem to have connections to various languages of apology, truthful, reconciliation, and promising: “truth believe lie true lying telling story,” “sorry life wasn wrong fault done love,” and “leave won away stay Please care promise.”

Unlike the streamgraph example, this removes time as an organizing feature of the analysis. Rather than use time, we can make two axes that correspond to the percentage of all words that are drawn from either of these two topics, and label from 1 to 12 the chunks of text moving through this space. Doing so shows an aggregate feature of plots: they move from high spatiality and low truthfulness to low spatiality and high truthfulness, with notable departures into other kinds of language in the first and last chunks (Fig. 4). Neither the x- nor the y-axis here is showing time directly; instead, the roughly linear flow is a manifestation of the intrinsic temporal organization of scripts.

Individual shows can be plotted into this same space. Limiting to 6 shows randomly sampled from the 20 most frequent shows in the database (Fig. 5), it is clear that these trends are present but not dominant at the scale of several dozen episodes. The two crime procedurals in the set, CSI and NCIS, show the aggregate trend most strongly. This is most likely because their rigidly procedural storytelling requires a crime to occur in the first half with a particular location, and then requires testimony, often in a courtroom, at the end. Other genres, however, including the science fiction show “Stargate SG-1” and, to some degree, the reality show “Survivor”, show the trends as well. The other two shows, the 2000s medical drama “Grey’s Anatomy” and the 1960s thriller “The Fugitive,” depart primarily in their opening or closing scenes; “The Fugitive,” the major plot point of which is Richard Kimble seeking exoneration for a crime he did not commit, opens with a tremendously large amount of “truth and reconciliation” language thanks in large part to the regular narration over the credits. The medical drama, on the other hand, departs primarily in its final scene. Although “Grey’s anatomy” does include a large quantity of apology language typical of most dramas, it simultaneously contains more an atypically large quantity of workplace language.
A. Principal Components in plot space

The example above is an arbitrary combination of 6 topics into two dimensions, not a general solution to describing plot movements. A more general characterization of the full universe of television plots uses all 127 topics and does not force any to exist on the same axis. Instead, each sixth of a script occupies a point in 127 dimensional space.

Although such a characterization is not possible to visualize in its full detail, the full space can be reduced down to two dimensions for inspection using Principal Components Analysis. The general plot arc is then projected into the space defined by the first two principal components, and the most significant loadings are projected into the same space so that their dimensionality can be compared in the same metric (Fig. 6).

This two-dimensional projection is essentially a reduction along the two phenomena shown above in figures one and two. There it was observed that many topics show a sustained pattern of linear directionality over the course of a script; many others are roughly the same at the beginning and end, but peak or trough in the middle. Since these two patterns are orthogonal to each other in time, the principal components analysis roughly places them on the two most important components. The first principal component is highly correlated with positive or negative directionality in time: the second (here displayed as the x axis) shows topics abnormally present or absent in the middle sections.

These arcs, in other words, can give a general formulaic expression to the characteristic feature of scripts that they have a beginning, middle, and end each of which carries certain elements in common with the other two. Some schools of narrative craft argue that stories trace circles where the characters return to their original situation. [11] These arcs can be seen as partially affirming that view—beginnings and ends hold important commonalities that middles lack. At the same time, the placement of directional change onto the first principal component emphasizes the substantial differences between the beginning and the ends of scripts.

The strength of these associations is more important than the shape. It is not surprising at all that PCA pulls out arc-shaped patterns; this is the characteristic “horseshoe effect,” and can be seen even in random walk data of the same size and shape. [12] There is sound reason, however, based on the underlying data to believe that the underlying arc shape exists; in essence, principal components analysis is doing the same work as the linear regression above to discriminate by two major dimensions. The amount of the variance explained by the model is substantially higher than random walk: when reducing a 12 by 100 matrix of 12ths of the script and the top 100 components, the first two of 12 principal components explain 93% of the variance. Comparing this to random walk data shows that the second component is in some ways the most remarkable one; there is a far greater amount (9.7 standard deviations) of remaining variance explained by it than in any of 1000 random simulations.

IV. CONCLUSIONS

The general arcs described by this method are not particularly surprising. Most watchers of television could easily report that scripts move from crime to trial or from conflict to resolution; others might well be aware that settings change from familiar to abstract, although the precise vocabulary for describing such a change is not entirely clear.

The potential usefulness of this method lies in providing a general vocabulary and framework for exploring differences in plots. It is at once commonsensical—relying on the various similarities and differences between beginning, middle, and end—and rooted entirely in the context of the narrative structures themselves. Jockers’ [1] strategy relies on a precomputed set to determine sentiment but allows a variety of plot structures to result from it; Piper’s [3] method, on the other hand, is fully
responsive to the language used in each text, but requires an initial intuition about a certain plot archetype. Plot arcs as described here combine the strengths of each; they do not presuppose a particular direction to the overall plot, but take their cue from the language of the plot itself.

In so doing, the vocabulary makes it possible to pose new questions. Is it possible to characterize several different fundamental plots by clustering the paths that scripts take in topic space? Preliminary experiments suggest some potentially fruitful avenues, but a 127 topic space may be too large for a show of only a few thousands words to be reliably characterized. What similarities in plot directionality can occur across genres? How does structure in political speech unfold? A suitably reduced vector space model of plot movements can help to provide a vocabulary for exploring questions like in humanities scholarship.

Fig. 7: Percentage of remaining variance explained by each Principal Component compared to random walk data

REFERENCES