Literature Fingerprinting: A New Method for Visual Literary Analysis

Daniel A. Keim*
University of Konstanz

Daniela Oelke†
University of Konstanz

ABSTRACT

In computer-based literary analysis different types of features are used to characterize a text. Usually, only a single feature value or vector is calculated for the whole text. In this paper, we combine automatic literature analysis methods with an effective visualization technique to analyze the behavior of the feature values across the text. For an interactive visual analysis, we calculate a sequence of feature values per text and present them to the user as a characteristic fingerprint. The feature values may be calculated on different hierarchy levels, allowing the analysis to be done on different resolution levels. A case study shows several successful applications of our new method to known literature problems and demonstrates the advantage of our new visual literature fingerprinting.

Keywords: Visual literature analysis, visual analytics, literature fingerprinting

Index Terms: J.5 [Computer Applications]: Arts and Humanities—Linguistics, Literature; I.6.9 [Visualization]: Information Visualization—Visualization Techniques and Methodologies

1 INTRODUCTION

Traditional literary analysis is mostly done without the use of a computer. One of the reasons is obvious: to properly understand a text not only the words that are used are important but also the context they are used in, and understanding them in the context is difficult to achieve algorithmically. However, there are some fields of literary analysis in which computers have already proved useful in the past. This includes the classification of texts and some aspects of literary criticism. Often these methods are based on features that are supposed to characterize the text. Feature extraction methods can be as simple as calculating the average sentence length or as complicated as estimating the vocabulary richness of a text. In the case of text classification often conventional classification methods such as Support Vector Machines or Bayesian Networks are used, that work fully automatically. In other applications, for example in the case of authorship attribution, more transparency is required. Then, nearest neighbor classification is a popular approach in which an unclassified document is attributed to the author with the most similar features given some reference documents with known authorship. For methods that are based on multidimensional feature vectors often a transformation to a low dimensional space is done (using PCA, SVD or Karhunen-Loève transform) and the results are visualized in a two-dimensional scatterplot [3, 9, 12] or with barcharts [9]. All the approaches have in common, that a single feature vector or value is used to characterize the whole text. This means that a lot of information is disregarded, since the change of the values as the text proceeds can reveal characteristic traits of an author or show interesting patterns (see fig. 1 for a first impression).

*e-mail: keim@inf.uni-konstanz.de
†e-mail: oelke@inf.uni-konstanz.de

IEEE Symposium on Visual Analytics Science and Technology 2007
October 30 - November 1, Sacramento, CA, USA
978-1-4244-1659-2/07/$25.00 ©2007 IEEE

Figure 1: Visualization of the two novels “The Iron Heel” and “Children of the Frost” by Jack London. Color is mapped to vocabulary richness. It can easily be seen that the structure of the two novels is very different with respect to this measure. This would be camouflaged if only a single value for each book would be calculated.

Our idea presented in this paper is to calculate the features for different hierarchy levels of the text (such as words, sentences, chapters, ...) and create a characteristic fingerprint of the text which contains significantly more information than just a single number and therefore enables the user to gain a deeper understanding. We successfully apply the method to known literature problems (e.g. authorship attribution) and show that by combining the automatic analytical methods of literature science with an effective visualization technique new insights into literary texts can be gained.

Outline

The rest of the paper is organized as follows: In section 2 the different types of computer-based literary analysis are introduced and different variables for literary analysis are briefly reviewed. Using some of the variables, in section 3 we then test their discrimination power with respect to the authorship attribution problem using novels of Mark Twain and Jack London. In section 4, we locally analyze the literature fingerprints of two novels of Twain and London and of a much bigger and more diverse text - the bible. Section 5 introduces our framework and finally, section 6 concludes the paper and outlines some interesting future applications of our new technique that are planned together with literature scientists.

2 BASICS FROM LITERATURE SCIENCE

2.1 Computer-based literary analysis

In a number of recent digital library projects, huge amounts of literature have already been digitized. In order to be able to computationally support the analysis of these texts algorithms are needed that can cope with all levels of natural language, namely the lexical, syntactic, and semantic aspects. Although the field of natural language processing has made significant progress over the last years, there are still a number of aspects the algorithms cannot cope with properly. This is especially true if the semantics or meaning of a text has to be taken into account, because of the vast amount of words, which have different meanings in different contexts, and the impressive flexibility and complexity of natural language. Still, computers are of great help whenever the lexical or syntactic structure of a text needs to be analyzed. In this section, we introduce two fields of computer-based literary analysis, in

which computers have already been successfully applied, namely text classification and different types of literary criticism.

Three different types of **text classification** can be distinguished: topic-oriented classification, classification into genres, and authorship attribution. For topic-oriented classification often TF-IDF vectors are used to characterize a text. TF-IDF (term frequency - inverted document frequency) vectors are made up of the frequency of each term in the document corpus weighted by the importance of that term with respect to the other documents in the corpus. Intuitively, a term is seen as characteristic for a document if its frequency within the document is much higher than its frequency in the rest of the corpus [10]. A typical application of topic-oriented classification is, e.g., labeling newspaper articles as Politics, Business, Sports, or Entertainment. For discrimination between different literary genres, such as Fiction / Non-Fiction, Children's literature, Travel literature, Poetry, etc., it is useful to consider the grammatical structure, the parts of speech used, and even the layout of the text in addition to TF-IDF vectors. In contrast, a special requirement of authorship attribution is that the extracted features should not be consciously controllable by the writer to prevent the method from being misdirected by a forged text. Note that for all mentioned methods the quality of the classification highly depends on the suitability of the features for discriminating the objects of the given categories. Therefore, enabling the user to understand the discrimination power of the features with respect to the classification task is of high importance.

Computer-assisted literary criticism is a rather young field in the studies of literature. According to [6], a frequently mentioned objection is that the words and sentences of a text cannot be analyzed without properly taking the context they are used in into account. Therefore, most researchers in literary studies only use computers to collect data that is afterwards analyzed conventionally. Yet, there are some cases in which the computer has already proven useful, e.g., for comparing an author's revisions (from version to version) or for the analysis of prosody and poetic phonology. Computer-assisted studies have also been performed in the context of sequence analysis, such as assigning quoted passages to speakers and locating them in the sequence of the text [6]. Another interesting field for computer-assisted analysis is translation criticism, in which metrics, rhythm, style, and other variables of the original text and the translation are compared to evaluate the quality of the translation.

All mentioned approaches have in common that one feature vector or value is calculated per text or per text block. Even if the text is split into chapters and paragraphs, usually the value for each chapter or paragraph is considered as a single object. In most cases, the values are averaged over the whole text, which leads to a smoothing of passages with an unusual trend, camouflaging interesting patterns. None of the computer-based literature analysis methods used so far deals with the behavior of the values across the text, which means that this very important information is completely ignored. In our approach, we therefore consider the literature in more detail by analyzing the texts on different hierarchy levels (i.e. calculating one value per sentence, paragraph, chapter, or text block). By visualizing the results of the detailed literature analysis together with their position in the text, even local analyses become possible. Moreover, the comparison of the visualizations for different variables leads to insights about the discrimination power of the different literature analysis variables. Since the success of each of the methods highly depends on an appropriate choice of the feature analysis variables, the possibility to efficiently compare the effectiveness of the variables with respect to a specific task provides new ways of an in-depth literature analysis.

2.2 Variables for literary analysis

Different variables for literary analysis have been proposed. They can roughly be classified into three groups: statistical measures, vocabulary measures, and syntax measures. In [8], a comprehensive survey on variables for literary analysis with a focus on authorship attribution can be found. Information about variables for text theme classification can be, for example, found in [7]. In this subsection, we briefly introduce some important text analysis measures to give the reader an overview of the field and provide the necessary background knowledge for the following sections. The focus will be on variables which measure the stylistic traits of literary texts in general and the style of an author in particular.

Statistical measures
Calculating the average *word length* or the average number of *syllables per word* are two simple variables to characterize a text. While the first one does not provide reliable results, the second one can be useful to distinguish different genres. This is intuitively plausible because in poetic texts the number of syllables of a word is much more important than in prose texts.
Sentence length is an indicator of style that can be used to estimate how good the rhythm of a text is preserved in a translation of the text. It is also used for authorship attribution studies, although in the context of authorship attribution it can be problematic since the length of the sentences is consciously controllable by an author and is not meaningful if the text has been edited by someone else. It has been shown that the distribution of sentence length is a more reliable marker for authorship than the average sentence length. Yet, it is also more difficult to evaluate. Here our technique proves useful because the visualization of the results allows an effective comparison of the distribution.
Instead of working on the words directly, it is also possible to analyze the proportions of certain *parts of speech (POS)* (such as nouns, verbs, adjectives ...) in the text. By this, the degree of formality of a text can be measured or the style of a text can be compared to its translation in another language.

Vocabulary measures
Vocabulary measures are based on the assumption that authors (and their texts) differ from each other with respect to vocabulary richness (how many words are in the vocabulary of the author and is s/he able to use his/her vocabulary by applying new words as the text proceeds) and with respect to word usage (which words are preferred if several can be applied).

To measure the characteristic word usage of an author the *frequencies of specific words* are counted. The success of this method highly depends on the appropriate choice of words for which the frequencies are compared. Different approaches have been suggested, e.g., to group the words into categories such as idiomatic expressions, scientific terminology, or formal words, and count the number of occurrences for each group or compare the frequency distributions of the words. Good results have been reported for function words such as "the, and, to, of, in ..." as the set of words. According to [5], function words have the advantage that writers cannot avoid using them, which means that they can be found in every text and almost every sentence. Furthermore, they have little semantic meaning and are therefore among the words that are least dependent on context. With the exception of auxiliary words they are also not inflected, which simplifies counting them. Finally, the choice of specific function words is mainly done unconsciously which means that it is an interesting measure for authorship attribution.

Measures of vocabulary richness are mainly based on the evaluation of the number of tokens and different types. In the following,

let N denote the number of tokens (that is the number of word occurrences which form the sample text, i.e. the text length), V the types (the number of lexical units which form the vocabulary in the sample, i.e. the number of different words), and V_r the number of lexical units that occur exactly r times. A simple measure for vocabulary richness is the *type-token ratio (R)* defined as

$$R = \frac{V}{N}.$$

This measure has one severe disadvantage, namely its dependency on the length of the text. A more sophisticated method to measure vocabulary richness is the *Simpson's Index (D)* that calculates the probability that two arbitrarily chosen words will belong to the same type. D is calculated by dividing the total number of identical pairs by the number of all possible pairs:

$$D = \frac{\sum_{r=1}^{\infty} r(r-1)V_r}{N(N-1)}.$$

While the Simpson's Index takes the complete frequency profile into account, there are also measures that focus on just one specific part of the profile. For example, [8] reports that Honoré suggested a measure that tests the tendency of an author to choose between a word used previously or utilizing a new word instead, which can be calculated as

$$R = \frac{100logN}{1 - V_1/V}$$

and is based on the number of *Hapax Legomena (V_1)* of a text, that means the number of words that exactly occur once. The method is said to be stable for texts with $N > 1300$. Similar to this, the *Hapax Dislegomena (V_2)* (the words that occur exactly twice) can be used to characterize the style of an author. According to [8], Sichel found that the proportion of hapax dislegomena (V_2/V) is stable for a particular author for $1,000 < N < 400,000$. At first this seems counterintuitively but with increasing text length not only more words appear twice but also words that formerly occurred twice now occur three times and therefore left the set of hapax dislegomena.
Many other methods to measure the vocabulary richness exist. The interested reader should consult [8] for a deeper investigation of the topic.

Syntax measures
Syntax-based measures analyze the syntactical structure of the text and are based on the syntax tree of the sentences. As the syntactical structure contains additional information, syntax measures have a high potential in literature analysis and have already been used in some projects. In [4], an experiment is reported in which a new syntax-based approach was tested against some word-based methods and was shown to beat them. In another approach [11], the authors build up syntax trees and develop different methods to analyze the writing style, the syntax depth, and functional dependencies by evaluating the trees. Note that – to a certain extend – the usage of function words also takes the syntax into account, because some function words mark the beginning of subordinate clauses or connect main-clauses. They therefore allow inferences about the sentence structure without analyzing the syntax directly.

3 Authorship Attribution

3.1 The concept of authorship attribution

The goal of authorship attribution is to determine the authorship of a text when it is unknown by whom the text has been written or when the authorship is disputed. Authorship attribution can also be used when there is doubt whether the person that claims to have written the text is really the creator. One example for such a doubtful situation is the assignment of the 15th book of the series of the Wizard of Oz. The book was published after the death of its author L. Frank Baum and was said to have been only edited by his successor Ruth Thompson who wrote the next books of the series. However, some literature specialists think that Ruth Thompson also wrote the 15th book and that the attribution to Baum was only due to commercial motives to ease the transition from one author to the next without losing sales. See [5] for an interesting analysis on the problem.

Authorship attribution has also been named stylometry, because the classification is based on the distinct stylistic traits of a document and is independent of its semantic meaning. To measure style, certain features of the text are extracted that clearly discriminate the literary work of one author from another author. Classical authorship attribution is mostly done on a pure statistical basis, excluding non-numeric measures. To get reliable results, enough texts of the potential writers with known authorship have to be available as basis for attributing the text in doubt to one of them.

3.2 Case study with literature of Mark Twain and Jack London

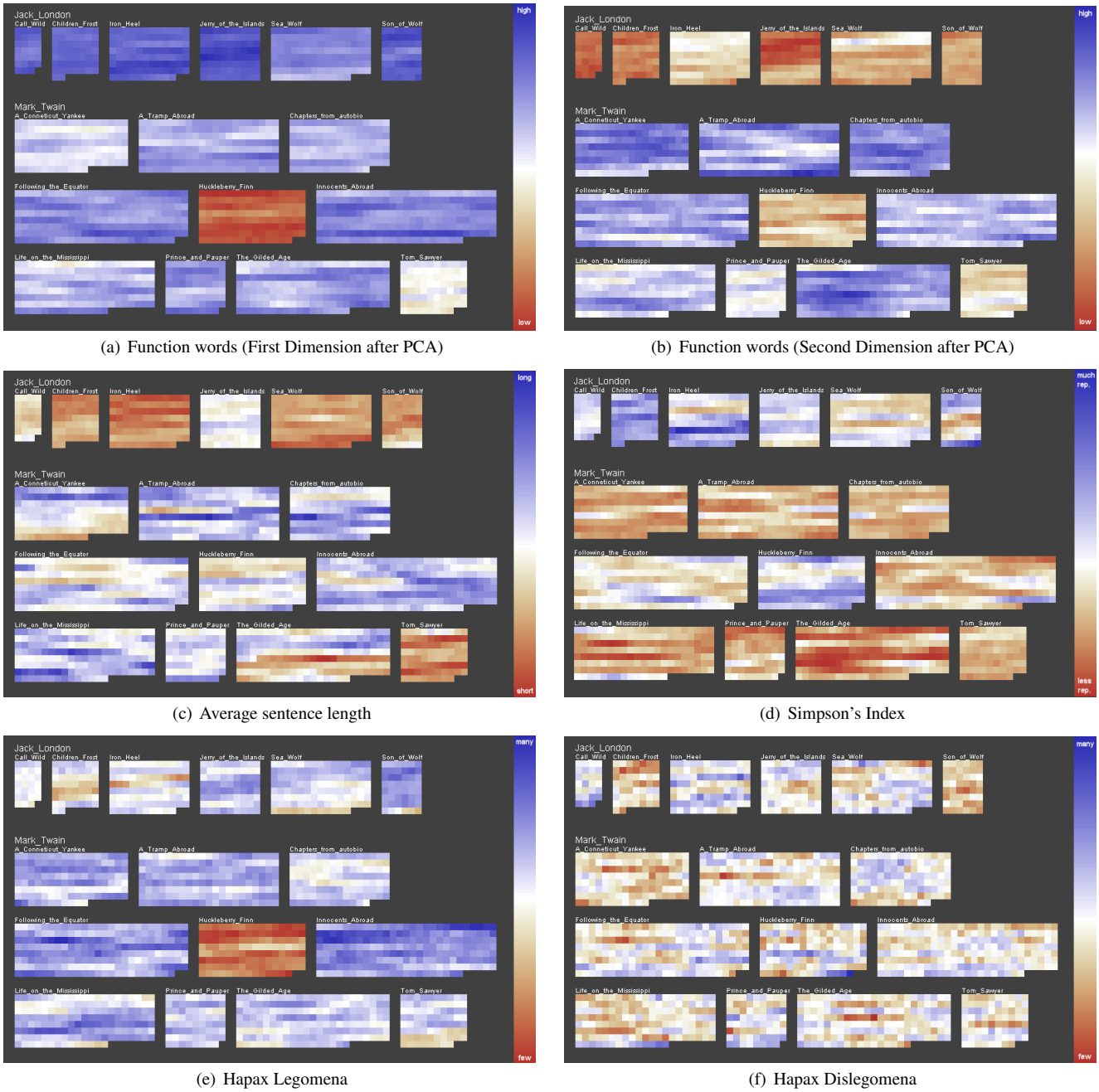
In this subsection, we will present the results of a study with literature of Mark Twain and Jack London. Our goal was to test the existing literature analysis measures and see whether our detailed visual representation leads to new insights.

In our study we used the following texts, that are all publicly available from Project Gutenberg [1]:

- Jack London:
  - *The Call of the Wild*
  - *Children of the Frost*
  - *The Iron Heel*
  - *Jerry of the Islands*
  - *The Sea-Wolf*
  - *The Son of the Wolf.*

- Mark Twain:
  - *A Connecticut Yankee in King Arthur's Court*
  - *A Tramp Abroad*
  - *Chapters From My Autobiography*
  - *Following the Equator*
  - *The Adventures of Huckleberry Finn*
  - *The Innocents Abroad*
  - *Life on the Mississippi*
  - *The Prince and the Pauper*
  - *The Gilded Age: A Tale of Today*
  - *The Adventures of Tom Sawyer.*

We preprocessed the texts by removing the preamble and other Gutenberg specific parts of the document and by replacing short forms with the corresponding long forms (e.g. isn't → is not). Afterwards we used the Standford POS tagger to annotate the texts [2]. For that we had to remove the chapter titles, since the tagger is only able to cope with complete sentences (though it is fault-tolerant with some grammatical errors). Finally, we split the documents into blocks with a fixed number of words each to be able to show the behavior of the variable values across the text. The number of words per block can be chosen by the user. For this paper, we set the number of words per block to 10,000, but similar results are obtained for a wide variation of this number as long as the blocks are not too small ($> 1,000$); since some literature analysis measures will provide unstable results when applied to short texts. To obtain a continuous and split-point independent series of values, we overlap the blocks with the neighboring blocks by about 9,000 words.



(a) Function words (First Dimension after PCA)

(b) Function words (Second Dimension after PCA)

(c) Average sentence length

(d) Simpson's Index

(e) Hapax Legomena

(f) Hapax Dislegomena

Figure 2: Fingerprints of books of Mark Twain and Jack London. Different measures for authorship attribution are tested. If a measure is able to discriminate between the two authors, the visualizations of the books that are written by the same author will equal each other more than the visualizations of books written by different authors. It can easily be seen that this is not true for every measure (e.g. Hapax Dislegomena). Furthermore, it is interesting to observe that the book *Huckleberry Finn* sticks out in a number of measures as if it is not written by Mark Twain.

This results in a soft blending of the values instead of hard cuts and therefore enables the user to easily follow the development of the values across the text.

As visual representation of the results we depict each text block as a colored square and line them up from left to right and top to bottom. Although very simple this is an effective visualization since the order of the text blocks is very important and the alignment corresponds to the standard reading direction. We also experimented with other shapes such as rounded rectangles, squares with beveled borders and circles. However, it turned out that the perception of a trend is easiest when displayed on a closed area with no borders visible. For the comparison of discrete values the other shapes are more useful. If a hierarchy has been defined on the text (made up of chapters, pages of the book, paragraphs, etc.), the pixels are visually grouped according to that hierarchy. Thereby, the structure of the text can be visually perceived and patterns that discern one passage of the other become obvious.

Since function word analysis is known as one of the most suc-

cessful methods for discriminating the texts of different authors, we started our analysis with this measure. We took a list of 52 function words that was also used in [5]. For each text block, a feature vector was calculated by counting the frequency of each of the function words, resulting in a 52-dimensional feature vector. We then applied principal component analysis (PCA) to the feature vectors to linearly transform the data to a new coordinate system in which the first dimension accounts for the largest variance, the second dimension for the second largest variance and so on. Figure 2(a) shows the values of the first dimension. We use a bipolar, interactively adjustable colormap to map the values to color. If a measure is able to discriminate the two authors, the books of one author will be mainly in blue and the books of the other one will be mainly in red. It is obvious that this is not the case here. What sticks out immediately is Mark Twains *The Adventures of Huckleberry Finn*. This novel seems to differ more from all the other writings of Mark Twain than the writings of the two authors differ from each other. If we visualize the second dimension of the transformed function word vectors we can see that the books of the two authors now separate from each other (figure 2(b)) - again with the exception of *Huckleberry Finn* (and this time also the book *The Adventures of Tom Sawyer*) which we would rather attribute to London than to Twain if its authorship was unknown. To analyze the strange behavior of *Huckleberry Finn*, we tested other variables such as Sentence length, Simpson's Index, the Hapax Legomena measure of Honoré, and the Hapax Dislegoma ratio (see section 2.2 for an introduction of the variables). Figures 2(c) - 2(f) show the visualizations for the different measures. In fig. 2(e) *Huckleberry Finn* again clearly stands apart. The Simpson's Index shown in fig. 2(d) would again mislead us to attribute the book to Jack London, whereas in 2(c) it nicely fits to all the other books of Mark Twain. Finally, the Hapax Dislegoma shown in 2(f) seems to have no discriminative power and is therefore not useful for the analysis. Taking all analysis measures into account, it is clear that there is something special about Mark Twain's *The Adventures of Huckleberry Finn*. The reasons for the exceptional behaviour cannot be answered by our analysis. The potential explanations range from language particularities such as the southern accent of the novel which may irritate some of the measures over the editing of the text in Project Gutenberg to the surprising speculation that a ghost writer was involved in the creating of the novel.

On the more general side, the figures show that not every variable is able to discriminate between the books of Mark Twain and those of Jack London, and this is also true if the novel *Huckleberry Finn* is excluded from the study. In fig. 2(f) (Hapax Dislegomena), we do not see much of a difference between the texts at all. The statement of Sichel that the proportion of Hapax Dislegomena in a text is specific for an author [8] cannot be verified, at least for these two authors. Instead, the sentence length measure (see fig. 2(c)) allows a very nice discrimination between the two authors. Mark Twain's books in average have longer sentences than Jack London's books. Only one novel per writer, namely *Jerry of the Islands* of Jack London and *The Adventures of Tom Sawyer* of Mark Twain break ranks and may be attributed to the other author. The second PCA dimension of the function word vector (fig. 2(b)) and the Simpson's Index (fig. 2(d)) also provide very nice results. Based on the Simpson's Index, we can observe a trend to a higher vocabulary richness (less repetition) in the writings of Mark Twain than in the books of Jack London.

4 Detail Analysis of Literature Fingerprints

The task of authorship attribution is an interesting application of our visual literature fingerprint. To reveal their full power, in this section we will look at the visual fingerprints in more detail.

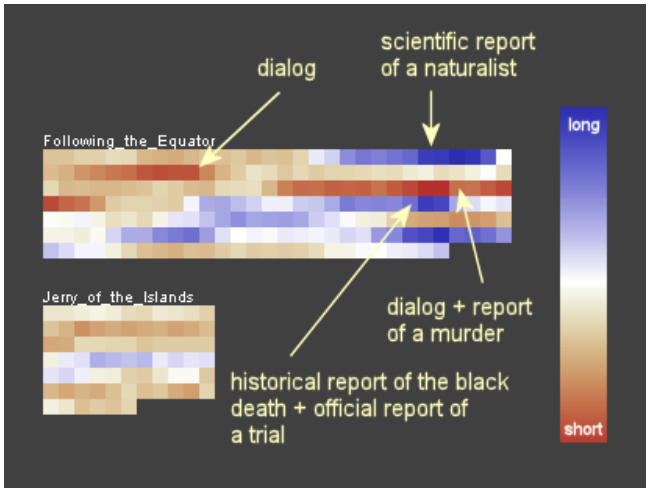


Figure 3: The figure shows the fingerprints of two novels that almost have the same average sentence length. In the detailed view, the different structure of the two novels is revealed. The inhomogeneity of the travelogue *Following the Equator* can be explained with the alternation of dialogs, narrative parts and quoted documents.

4.1 Detail analysis of two novels

In this subsection, we will analyze two books, whose average sentence length is about the same. The images in Figure 3 show the result of splitting the text into overlapping text blocks of 10,000 words each (with an overlap of 9,000 words) and calculating the average sentence length block-wise. The visual fingerprints reveal that the structure of the two books is totally different despite their identical overall average values. While the average sentence length in *Jerry of the Islands* of Jack London does not differ much across the novel (and thus the total average value would be meaningful), there are significant variations in *Following the Equator* of Mark Twain. *Following the Equator* is a non-fiction travelogue that Mark Twain wrote as an account of his tour of the British Empire in 1895. In fig. 3, some passages stick out as they are in dark blue respectively dark red. Taking a closer look at the text reveals the reasons: The long stripe in dark blue in the first line, for example, represents a passage, in which Mark Twain quotes the scientific text of a naturalist with rather complex and long sentences. On the other hand, in the dark red passages in the second and third line Mark Twain noted some conversations that he had during his travel with the short sentences of spoken language. The second dialog is directly followed by the quotation of a written report about a murder. One would rather expect such a report as being characterized by long sentences. This is probably why Twain himself utters his surprise about the text in his book. He says:

> *"It is a remarkable paper. For brevity, succinctness, and concentration, it is perhaps without its peer in the literature of murder. There are no waste words in it; there is no obtrusion of matter not pertinent to the occasion, nor any departure from the dispassionate tone proper to a formal business statement."* [13]

The dark blue area in the forth line is due to a historical report of the black death and an official report of the trail.

4.2 Detail analysis of the bible

In a second study, we analyzed the visual fingerprint of the bible. In this case, we used the existing hierarchy of the text to define the blocks. While every text has an inherent syntactical

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.