Literature Fingerprinting:
A New Method for Visual Literary Analysis
(based on the paper by Daniel A. Keim and Daniela Oelke)

Simon Wegener
Universität des Saarlandes

29.05.2008
A Little Example

Figure 5: Visual Fingerprint of the Bible. More detailed view on the bible in which each pixel represents a single verse and verses are grouped to chapters. Color is again mapped to verse length. The detailed view reveals some interesting patterns that are camouflaged in the averaged version of fig. 4.

Figure 6: The visualization of this paper helped to find the longest sentences and improve the readability by modifying them appropriately.
Outline

Introduction

Local Literary Analysis

Authorship Attribution

Conclusion
Literary Analysis

- What is literary analysis?
- It is ...
  - study of structure,
  - classification,
  - interpretation,
  - critique
  ... of a text in the context of ...
  - genre,
  - author,
  - history.
Literary Analysis and the Computer

- Is the computer a good literary analyst?
- Unfortunately, not really. Why?
- Computers can’t figure out the semantics of human language very well. To properly understand a text, the context of the words in the text is important. But understanding the words in their context is very difficult to achieve algorithmically.
Computer-Assisted Literary Analysis

- How can we benefit from the computer?
- Use the computer for what it was made for: basic math.
- Counting lines, pages, words.
- Calculate the average sentence length.
- Compute complex measures.
The Classic Approach in Computer-Assisted Literary Analysis

1. Compute a feature vector for the whole corpus (i.e. text).
2. Use this single value for conventional literary analysis.
A Poem

Not how did he die, but how did he live.
Not what did he gain, but what did he give.
These are the units to measure the worth
of a man as a man, regardless of birth.
— anonymous
Possible Analysis Output

Analysis found 37 words at 4 lines.
Average sentence length is 12.333, average word length is 3.243.
The average syllable count per word is 1.054,
35 words with 1 syllable, 1 word with 2 syllables, 1 word with 3 syllables.
Type-token ratio is 0.621.
Found 13 hapax legomena ...
Tag Clouds
Literature Fingerprinting

1. Compute feature vectors for different hierarchy levels.
2. Visualize them in a characteristic fingerprint of the text.
3. Gain a deeper understanding. (Hopefully!)
Local Literary Analysis

• First case study: local analysis of two books by Jack London and Mark Twain with almost the same average sentence length
• Does the analysis benefit from the visual fingerprint?
Local Literary Analysis

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 section we will look at the visual fingerprints in more detail.

Following the Equator

Jerry of the Islands

The Adventures of Tom Sawyer

The Adventures of Huckleberry Finn

Huckleberry Finn

The reasons of Jack London does not differ much across a formal business statement.

For brevity, succinctness, and nature of murder. There are no waste words in it; there

It is a remarkable paper. For brevity, succinctness, and

Following the Equator can be explained with the

can be explained with the

may be attributed to the other author. The second PCA dimension 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 us to discriminate between the two authors. Mark Twain’s books in average have longer sentences than Jack London’s books. This novel seems to differ more from all the other writings of Mark Twain if its authorship was unknown. To analyze the strange behavior cannot be answered by our analysis.

On the more general side, the figures show that not every variable measures over the editing of the text in Project Gutenberg to the sureties into account, it is clear that there is something special about the dark blue area in the forth line is due to a historical report of a murder. 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:

In fig. 2(f) (Hapax Dislegoma) seems to have no discriminative power for the exceptional behavior cannot be answered by our analysis.

The Simpson’s Index shown in fig. 2(d) would again clearly

The long stripe in dark blue in the first line, for example, represents a passage, in which Mark Twain quotes the scientific text of Twain wrote as an account of his tour of the British Empire in 1895. There are significant variations in

identical overall average values. While the average sentence length is about the same. The images in Figure 3 show the

The dark blue area in the forth line is due to a historical report of
twenty words each (with an overlap of 9,000 words) and calculating the result of splitting the text into overlapping text blocks of 10,000 words, resulting in a 52-dimensional feature vector. We then applied principal component analysis (PCA) to the feature vectors to

For the exceptional behavior cannot be answered by our analysis.
Authorship Attribution

• Determine, if a text was written by an author or not.
• Is a common problem in literary analysis.
• What features are useful for discrimination?
• Case study on some books by Jack London and Mark Twain.
Variables for Literary Analysis

- **Statistical measures**
  - Syllables per word
  - Sentence length
  - Proportions of parts of speech
  - ...

- **Vocabulary measures**
  - Frequencies of specific words
  - Type-token ratio
  - Simpson’s index
  - Hapax (dis)legomena
  - ...

- **Syntax measures**
Hapax Legomena

• A hapax legomenon is a token, that occurs only once in the corpus.
• The example poem contains 13 hapax legomena.

Not how did he die, but how did he live.
Not what did he gain, but what did he give.
These are the units to measure the worth of a man as a man, regardless of birth.
— anonymous
Hapax Dislegomena

- A hapax dislegomenon is a token, which occurs exactly twice in the corpus.
- The example poem contains 8 hapax dislegomena.
- Is said to be able to characterize the style of an author.

\textit{Not how did he die, but how did he live.}
\textit{Not what did he gain, but what did he give.}
\textit{These are the units to measure the worth of a man as a man, regardless of birth.}
— anonymous
Definitions

- $N \overset{\text{def}}{=} \text{number of tokens in the corpus (i.e. text length)}$
- $V \overset{\text{def}}{=} \text{number of types in the corpus (i.e. different words)}$
- $V_r \overset{\text{def}}{=} \text{number of types that occur exactly } r\text{-times}$
99 Bottles of Beer

99 bottles of beer on the wall, 99 bottles of beer.  
Take one down and pass it around, 98 bottles of beer on the wall.

98 bottles of beer on the wall, 98 bottles of beer.  
Take one down and pass it around, 97 bottles of beer on the wall.

...

1 bottle of beer on the wall, 1 bottle of beer.  
Take one down and pass it around, no more bottles of beer on the wall.

No more bottles of beer on the wall, no more bottles of beer.  
Go to the store and buy some more, 99 bottles of beer on the wall.
Type-Token Ratio

\[ R = \frac{V}{N} \]

- A simple measure for vocabulary richness.
- The type-token ratio of the example poem is \( \frac{23}{37} = 0.62162 \)
- The type-token ratio of „99 bottles of beer“ is 0.00992.
Simpson’s Index

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

- A more sophisticated measure for vocabulary richness.
- Describes the probability that two tokens belong to the same type.
- The Simpson’s index of the example poem is \( \frac{1 \cdot 0.13 + 2 \cdot 1.8 + 3 \cdot 2.0 + 4 \cdot 3.2}{37.36} = \frac{40}{1332} = 0.03003 \).
- In comparison, the Simpson’s index of „99 bottles of beer“ is 0.09307.
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 successful approaches...
Authorship Attribution

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 successful...
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 successful...
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 successful...
Authorship Attribution

(c) Average sentence length
Authorship Attribution

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 successful...
Summary

What have we seen so far?

- Some techniques of computer-assisted literary analysis.
- Some measure for literary analysis and their ability to discriminate between authors.
- A visualisation for computer-assisted literary analysis.
- Two successful applications of this method.
Further Applications

- Genre classification
- Translation criticism
- Formality analysis
- Readability analysis
Related Work

- Tag Clouds
- Word Trees
- SeeSoft
I am happy to join with you today in what will go down in history as the greatest demonstration for freedom. Not unmindful that some of you have come here out of great trials and tribulations.

Must say to my people who stand on the warm threshold which leads into the palace of justice.

Say to you today, my friends, so even though we face the difficulties of today and tomorrow, I still have a dream.

I still have a dream.

This nation will rise up and live out the true meaning of its creator's dream.

This nation will rise up and live out the true meaning of its creator's dream.

My four little children will one day live in a nation where they will not be judged by the color of their skin but by the content of their character.

Go back to the south with.
The Tool

Figure 7: Screenshot of the framework that we implemented to enable an efficient and interactive analysis of texts.
Fig. 2. Visual analytics integrates scientific disciplines to improve the division of labor between human and machine.

Definition of Visual Analytics

In “Illuminating the Path” [39], Thomas and Cook define visual analytics as the science of analytical reasoning facilitated by interactive visual interfaces. In this chapter, however, we would like to give a more specific definition:

Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets.

The goal of visual analytics is the creation of tools and techniques to enable people to:

– Synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data.
– Detect the expected and discover the unexpected.
– Provide timely, defensible, and understandable assessments.
– Communicate assessment effectively for action.

By integrating selected science and technology from the above discussed disciplines and as illustrated in Figure 2, there is the promising opportunity to form the unique and productive field of visual analytics. Work in each of the participating areas focuses on different theoretical and practical aspects of users solving real-world problems using Information Technology in an effective and efficient way. These areas have in common similar scientific challenges and significant scientific added-value from establishing close collaboration can be identified. Benefit of collaboration between the fields is identified to be two-fold:

– Jointly tackling common problems will arrive at better results on the local level of each discipline, in a more efficient way.
– Integrating appropriate results from each of the disciplines will lay the fundament for significantly improved solutions in many important data analysis applications.
Critique

The Visualisation is ...

- easy to understand,
- good for comparisons,
- and shows new and interesting insights.

But:

- Only one feature at a time is visualized.
- Scaling?
The Bible

Figure 5: Visual Fingerprint of the Bible. More detailed view on the bible in which each pixel represents a single verse and verses are grouped to chapters. Color is again mapped to verse length. The detailed view reveals some interesting patterns that are camouflaged in the averaged version of fig. 4.

Figure 6: The visualization of this paper helped to find the longest sentences and improve the readability by modifying them appropriately.
Thanks!

Discussion.