Author Identification in Chatlogs using Formal Concept Analysis

Luuk van der Knaap           Franc Grootjen

NICI, Radboud University Nijmegen, the Netherlands

Abstract

In this project a new approach to author identification in chatlogs is investigated. The method is based on Formal Concept Analysis (FCA): a classification technique that is able to classify users of a chat-channel based on the words they posted. The concepts produced by FCA are scored using various evaluation functions.

The results of this explorative survey are most promising. Although identification is far from perfect and many tests still need to be done, FCA certainly has potential as a method for author identification in chat texts.

1 Introduction

Since the internet became a normal part of everyday life in most Western countries, the use of online chat-boxes and instant messengers has grown enormously, especially among teenagers. Unfortunately, the anonymity of the chat-box can easily be misused to, for example, persuade youngsters to cross borders they would never have crossed in a face-to-face situation. The anonymity of the internet accommodates forms of so-called ‘cyber crime’: the online business of criminal activities. The abuse of chat-channels, for whatever purpose, is simply one of them. Seeing that fighting online criminals, despite some successes, had led to a technological race between offenders and Justice in order to trace IP-addresses and their physical locations, we believe that generalizing over the entire range of internet crimes will not be the most successful course of action. It is our opinion that the various online activities differ so much that it will be more effective to develop specialist crime fighting tools for each of them.

When focussing on chatting only, there is a remarkable characteristic of this medium that catches the eye: the user is able to precisely control how much of his identity is published. The availability of this function has consequences for both user and misuser. For the user, it creates the sense of anonymity mentioned earlier. For the misuser, however, it first of all provides a means of misleading innocent users by presenting himself under a false and for the victim trustworthy identity. Further, it reduces the criminal’s risk to get caught, because he can create an entirely new identity for each crime. It is therefore our opinion that, although the technique presented in this thesis might affect the privacy of innocent chat-channel participants if it is misused, being able to link the chat-texts from various online personalities to a single author would be a great step forward in fighting potential criminals.

In this survey, we will attempt to identify authors from the chatlogs they produced while chatting on a public IRC-channel. Using Formal Concept Analysis (FCA) [17] we will map groups of authors with the same word usage into concepts. A scoring function will allocate points to authors based on their occurrence in concepts. This way we obtain a ranked list of authors which will be used for identification purposes.

2 Author Identification

“In the beginning, God created the heavens and the earth.”1 These words are not only the beginning of the most-read book in history and, according to this book, of the world as we know it, but have also raised a lot

1Gen. 1:1, KJV
of questions about who wrote them. At first, Moses was believed to have written the first book of the Bible, Genesis, from which this citation is taken, but later research shows that at least three authors contributed to this book [2, 16, 15, 18]. These and other surveys on the Bible’s authorship formed the beginning of a research field known as author identification.

Although analyzing the Bible is still a day-job for many researchers around the world, author identification has been applied in some more historical projects. In the 1960’s, Mosteller and Wallace performed their seminal study ‘The Federalist Papers’ in which 12 out of 146 political essays from the late eighteenth century were claimed by two authors: Alexander Hamilton and James Madison. Using a Bayesian statistical analysis to the frequencies of function words, Mosteller and Wallace were able to assign the correct author to all twelve papers [12]. The same result was achieved by Kjell [11], who made use of a neural network to assign the group of unknown documents as a whole to a single author. This success was not equalled by the other famous historical project: the ‘Claremond Shakespeare Authorship Clinic’, in which Ward Elliott and Robert Valenza tried to find the ‘true’ Shakespeare from among 37 possible authors by means of quantitative text analysis [8]. Two other researchers, Merriam and Matthews, where able to classify some texts by building a neural network that was able to distinguish between two authors: Shakespeare and Marlow [14].

2.1 Modern computational linguistics

A fairly complete overview of rather recent research on author identification in modern texts is given by McCombe [13], from who most of the information below is taken.

2.1.1 Chaski’s hypotheses

In linguistics, one of the leading authorities on author identification is Chaski. She states that “...the fact of language being an automatic process gives us some scope for linguistic fingerprinting: the more automatic a process is, the less control we have over it and the greater [is] its reliability as an indicator of individuality” [3].

In a later report, Chaski tests nine empirical hypotheses that have been used to identify authors in the past [4]. Not all of them were equally successful, and therefore it was Chaski’s intention to perform a critical survey. She examined the following measures: Vocabulary richness, Hapax Legomena (words occurring once), Readability measures, Content Analysis (semantic classification), Spelling errors, Grammatical errors, Syntactically classified punctuation, Sentential complexity, and Abstract syntactic structures. Chaski’s conclusion is that the only reliable methods that appear to accurately identify authors are those relying on linguistic science and generative grammar: the sentential complexity method and the method based on abstract syntactic structures.

2.1.2 The response by Grant and Baker

In [10], Grant and Baker respond to Chaski’s publications. First of all, they publish a short overview of the history of author identification in which they refer to the Shakespeare and Federalist projects and explain that some methods strictly rejected by Chaski have worked well in earlier and widely-acknowledged studies. Furthermore, they describe their own approach to the author identification problem, known as ‘Principal Component Analysis’, which “identifies which marker or combination of markers discriminates in a particular case as particularly effective” [10]. At first sight, this technique appears to be less broadly applicable in comparison to Chaski’s methods, but the authors explain that this is a misunderstanding. They emphasize that Chaski uses a deliberately limited sample database, in order to minimize the variation between her subjects. Although this appears to make the identification task harder (identical subjects produce identical texts, which are harder to distinguish from each other), it also has a negative effect on the generalisability of her conclusions to a wider population, in which variation between subjects is rather common [13].

2.2 Identification using topic-free representations

In their 2005-paper, Madigan et al. approach the identification problem in an entirely new way. They foreground that the most important part of the identification process is the representation of the corpus. In stead of all their preceding colleagues, they chose a topic-independent representation of the texts. In their
experiments, they decided to use a corpus containing a variety of topics and they acknowledge that it will be a difficult task to represent a corpus with a central theme in a topic-independent way.

After implementing their representation technique, Madigan et al. used a Bayesian multinomial logistic regression to classify the texts. This technique is able to cluster the texts with a certain probability or error rate. In their conclusions, Madigan et al. state that, although the initial experiments “...suggest that sparse Bayesian logistic regression coupled with high-dimensional document representations show considerable promise as a tool for authorship attribution [...] significant challenges concerning representation remain; different document representations can lead to different attributions and no clear method exists for accounting for this uncertainty” [12].

2.3 Author identification in e-mails

In the period 2001-2003, a series of articles and a thesis was published by de Vel, Anderson, Corney and Mohay. In their publications, they investigated the learning and classification of authorship categories in both aggregated and multi-topic e-mail documents. Their corpus was represented using a large set of content-free features, like structural characteristics and linguistic patterns. The documents were classified using the Support Vector Machine learning algorithm. For both aggregated and multi-topic e-mails, the results were encouraging. The authors themselves, however, prefer to experiment with a larger number of authors before calling the technique a success [7, 5, 6].

2.4 Research on author identification in chatlogs

When reviewing nearly all the currently published literature on author identification, it caught our attention that almost no research is specifically dedicated to author identification in chatlogs. This is remarkable, because the communication in chatlogs is on various dimensions (reduced grammar, abbreviations, mixed languages etc.) unique, which means that conventional identification techniques are not likely to work. It also means that, as far as we know, this is the first attempt to identify authors using a technique like FCA. That fact makes this project a truly explorative survey. Within our limited time frame, there is therefore no room left for exhaustive statistical analyses for each variable, although that is of course the first recommendation for future research.

3 Formal Concept Analysis

In FCA [17], a conceptual structure is derived from two sets and their binary relation. Within our domain we recognize a set of authors (objects) and the words they use (attributes). We denote the set of objects as $O$ while the set of attributes is represented by $A$. The result of indexing is a binary relation between objects and attributes. This relation is called a context and written as $\sim$. Note that $\sim \subseteq O \times A$.

Example 1

Consider the following (extremely short) chatlog:

john: hey guys, how are you?
mary: I am fine
pete: hey john, I am ok!

Indexing yields: $O = \{john, mary, pete\}$ and $A = \{hey, guys, how, are, you, I, am, fine, john, ok\}$. Furthermore it is easy to see that for example $john \sim guys$ and $pete \sim ok$. The complete context relation is presented in table 1.

<table>
<thead>
<tr>
<th></th>
<th>hey</th>
<th>guys</th>
<th>how</th>
<th>are</th>
<th>you</th>
<th>I</th>
<th>am</th>
<th>fine</th>
<th>john</th>
<th>ok</th>
</tr>
</thead>
<tbody>
<tr>
<td>john</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mary</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pete</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Example chatlog context
Let $o \in O$ be an object; $a \in A$ be an attribute. $O$ and $A$ represents subsets of respectively $O$ and $A$. First of all, we extend the context relation $\sim$ to sets:

$$a \sim O \equiv a \sim o \text{ for all } o \in O$$

$$o \sim A \equiv o \sim a \text{ for all } a \in A$$

Objects may share attributes. To express this notion we define the function $\text{ComAttr}$:

$$\text{ComAttr}(O) = \{ a \in A \mid a \sim O \}$$

Similarly attributes may share objects:

$$\text{ComObj}(A) = \{ o \in O \mid A \sim o \}$$

Although it seems plausible to assume that the functions $\text{ComAttr}$ and $\text{ComObj}$ are each others inverse, in general they are not. When they are, we have a special situation: A pair $(O, A)$ is called a (formal) concept if and only if:

$$\text{ComAttr}(O) = A$$

$$\text{ComObj}(A) = O$$

How to find all concepts for a given context falls outside the scope of this article. One of the most commonly used algorithms is developed by Ganter [9], which is adapted for parallel execution by Blokpoel [1]. The concepts derived from our example are presented in table 2.

<table>
<thead>
<tr>
<th>concept</th>
<th>objects</th>
<th>attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>∅</td>
<td>hey, how, are, you, guys, I, am, fine, john, ok</td>
</tr>
<tr>
<td>2</td>
<td>john</td>
<td>hey, how, are, you, guys</td>
</tr>
<tr>
<td>3</td>
<td>pete</td>
<td>hey, john, I, am, ok</td>
</tr>
<tr>
<td>4</td>
<td>pete, mary</td>
<td>I, am</td>
</tr>
<tr>
<td>5</td>
<td>pete, john</td>
<td>hey</td>
</tr>
<tr>
<td>6</td>
<td>pete, john, mary</td>
<td>∅</td>
</tr>
</tbody>
</table>

Table 2: All concepts in our example

Concepts 1 and 6 are special, they are called the top and bottom concepts of our structure. Furthermore we see that pete and mary are joined in concept 4 since they both used the words I and am. Same holds for pete and john (concept 5) using the word hey. We represent the set of all concepts with the symbol $C$.

4 The experiment

To test FCA for author identification we took the chatlog of a dutch, public IRC-channel. After removing some source code and other noise from the log file, there were approximately 25,000 lines of data left. For practical reasons, mostly due to memory limitations and calculation time, this file was split-up into 5 parts containing about 5,000 lines of utterances. To measure the effectiveness of the identification we picked out a user (later referred to as $X$), and changed his/her username to test in 20% of the utterances. The goal of the experiment was to determine if we can identify test being the same person as $X$. Finally we indexed the chatlog’s utterances, creating a context and the concepts for each part.

4.1 Scoring the concepts

In this section we describe three different ways to score the generated concepts in order to identify the test user. On average each chatlog part generated about 7000 different concepts. Some of them included both user test and $X$, some only $X$, some only test and some neither one of them. We are going to assign an Identification Score (IS) to each author using three strategies. We start with defining the occurrence function which expresses if an author $o$ co-occurs with test in a concept $(O, A)$.

$$\text{occur}(o, (O, A)) = \begin{cases} 1 & \text{if } o \in O \land \text{test} \in O \\ 0 & \text{otherwise} \end{cases}$$
4.1.1 Plain identification

Every time an author shows up in a concept in which the test user appears too, should be rewarded. So the plain score ($IS_p$) adds 1 point to the author when it appears in the same concept as the test user.

$$IS_p(o) = \sum_{c \in C} \text{occur}(o, c)$$

4.1.2 Attribute length identification

In the plain score, every concept containing both the author and the test user is equally important. However, concepts with a lot of attributes (shared words) may be of greater value for the identification process. That is why we suggest an alternative scoring mechanism, the attribute length score:

$$IS_a(o) = \sum_{(O, A) \in C} \#(A) \cdot \text{occur}(o, (O, A))$$

4.1.3 Object length identification

Another way to assign a higher score to important concepts is to pay attention to the number of objects in the concepts. If the number of objects is smaller, then the concept is more identifying for those users and should therefore receive a higher score. This leads to the object length score:

$$IS_o(o) = \sum_{(O, A) \in C} \frac{\text{occur}(o, (O, A))}{\#(O)}$$

5 Results

5.1 Plain identification scores

The results for scoring the users following the plain identification method are in figures 1 through 3. The average scores over all five parts can be found in figure 3. At first sight, the results give reason for optimism: the user that we are looking for, $X$, ends up in the top 5 scores of candidates in all parts except the first. In parts two and three $X$ even appears as the first candidate. On average, he reaches a second place over all five parts.

![Figure 1: Plain identification scores in the first and second part of the chatlog.](image-url)
5.2 Attribute length identification scores

The same way, we calculate the results for attribute and object length identification, from which we will not show the histograms here. Concerning attribute length identification, the first thing to notice is that the ranklist in table 3 barely changes. X is still the first candidate in two parts (2 & 3) and even dropped down one position in part 5. On average, his position does not change.

<table>
<thead>
<tr>
<th>score</th>
<th>part1</th>
<th>part2</th>
<th>part3</th>
<th>part4</th>
<th>part5</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS_p</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>IS_a</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3: The rank of X in all 5 parts of the chatlog

It is, however, difficult to compare these ranks to those of the plain identification scores since they do not provide information about the distance between X and his competitors. In table 4 we therefore determine for all tests the ratio between X’s IS and the average IS of the other users, and compare these numbers to those of the plain identification scores. These numbers indicate that the results do not improve enormously when comparing attribute length identification to plain identification. The ratio is on average increased by only 1%. Therefore, attribute length identification appears not to be a very useful scoring function for FCA.
5.3 Object length identification

Object length identification also does not much influence the results, as can be seen in table 5. $X$ is still the first candidate in two parts (2 & 3) and climbed one position in part 1. On average, his position does not change.

<table>
<thead>
<tr>
<th>score</th>
<th>part1</th>
<th>part2</th>
<th>part3</th>
<th>part4</th>
<th>part5</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>$IS_p$</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>$IS_a$</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>$IS_o$</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5: The rank of $X$ in all 5 parts of the chatlog

We calculate the same ratios as in attribute length identification (table 6). These results are slightly better than those of attribute length identification.

<table>
<thead>
<tr>
<th></th>
<th>part1</th>
<th>part2</th>
<th>part3</th>
<th>part4</th>
<th>part5</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>$IS_p$</td>
<td>997</td>
<td>2344</td>
<td>2084</td>
<td>1611</td>
<td>204</td>
<td>1448</td>
</tr>
<tr>
<td>$IS_a$</td>
<td>132.20</td>
<td>297.59</td>
<td>266.50</td>
<td>204.11</td>
<td>30.56</td>
<td>186.19</td>
</tr>
<tr>
<td>$IS_o$</td>
<td>88.74</td>
<td>107.39</td>
<td>89.46</td>
<td>75.77</td>
<td>20.72</td>
<td>55.27</td>
</tr>
<tr>
<td>ratio (plain)</td>
<td>1.453</td>
<td>2.617</td>
<td>2.715</td>
<td>2.541</td>
<td>1.496</td>
<td>3.195</td>
</tr>
<tr>
<td>ratio (object)</td>
<td>1.490</td>
<td>2.771</td>
<td>2.979</td>
<td>2.694</td>
<td>1.475</td>
<td>3.369</td>
</tr>
</tbody>
</table>

Table 6: Quantifying the quality of the effect of object length identification

6 Conclusions

In our first attempt, the plain identification scores (section 5.1), we were surprised by the performance of such an unrefined approach to IS-assignment. Being able to rate $X$ as the best-matching candidate in two out of five parts, and achieving a second place on average, emphasizes the great potential of this technique.

The effect of our two attempts to refine the assigned identification scores by assigning a weight to the concepts was also surprising: there barely was any effect at all. On average, performance increased by 1% (table 4, section 5.2) when the weight of the concept was derived from the length of it’s attribute-side. The third evaluation function, object length identification, improved performance by 5% on average (table 6, section 5.3). The usability of these increases is, of course rather low, but the results do show that performance can be improved using better evaluation functions and that it is therefore worth the effort to continue research on this matter.

As a general conclusion we can say that, given the promising results in this first survey, Formal Concept Analysis has a high potential as an automated identification technique in chatlogs. An increase in performance is likely to be gained by more experimentation with the evaluation functions.
References


