Mining Police Digital Archives to Link Criminal Styles with Offender Characteristics

Richard Bache[1], Fabio Crestani[1], David Canter[2], and Donna Youngs[2]

[1] Department of Computer and Information Science, University of Strathclyde, Scotland
{r.bache,f.crestani}@cis.strath.ac.uk
[2] Centre for Investigative Psychology, England
dcanter@ukonline.co.uk

Abstract. The partial success in inferring the characteristics of offenders from their criminal behaviour ('offender profiling') has relied on limited data and subjective judgments. We therefore sought to determine if Information Retrieval techniques and in particular Language Modelling could be applied directly to existing police digital records of criminal events to identify significant characteristics of offenders. The categories selected were gender and age group. Results showed that distinct differences in characteristics do exist.

Keywords: Document Classification, Text Data Mining, Language Models, Crime Data, Investigative Psychology, Offender Profiling.

Since the earliest criminological studies it has been clear that, broadly speaking, criminals have characteristics that distinguish them from the general population. There have also been attempts to demonstrate that certain classes of crime are typically committed by people who have similar characteristics. It has also been claimed that what may be called the 'style' of the crime, or the pattern of behaviour, typical of any set of crimes relates directly to subsets of characteristics of offenders. This process of making inferences about significant features of an offender on the basis of the kinds of people who commit crimes in that style has often been called 'offender profiling'. In general such 'profiles' are drawn from the subjective judgement and experience of putative experts with little empirical basis for their claims.

However, the few empirical studies that have been carried out (e.g. [1]) to develop models relating offence style to offender characteristics have relied on intensive content analysis procedures that derive categories from open-ended police and related data sources. Such procedures are both prone to subjectivity and require great human effort making them difficult for police officers to use in the field. However the emerging application of text mining of descriptions available in police digital records [2] provides technologies for performing such analysis automatically. But although it is reasonably straightforward to derive tokens in a systematic, objective fashion from police records, thereby mechanizing the development of the content categories there is still the empirical question of whether the tokens so derived do indeed provide the basis for discriminating between different categories of offender.

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 493–494, 2007.

The present study was therefore set up to establish whether two crucial characteristics of an offender, age and gender, could be reliably indicated using Language Models applied to actual police records, beyond the base rate levels of these characteristics in the offending population. Of course, in principle, this approach can be extended to many other characteristics. The work thus has three possible uses:

1. We can determine whether police records can be examined to reveal behavioural differences between categories of offender.
2. For an unsolved crime with no eye witness, the likely characteristics of the offender can be inferred and thus used to limit the range of possible offenders investigators should consider or to prioritise plausible suspects.
3. Specific features can be linked to offender characteristics. This can inform police investigating crimes as to the likely features to be associated with known categories of offender. This can assist for instance in evaluating witness testimony or hearsay evidence.

It can be argued that the problem addressed here is one of document classification and Probabilistic Language Modelling has been extensively used for this [3, 4, 5]. However, we argue that we are going beyond classification since we are firstly determining if behavioural differences exist between categories at all and secondly we are using the probabilities of terms in the language models to reveal behavioural styles in each group.

The rationale for being able to classify sex or age of offenders from their actions rests on there being significant behavioural differences between these groups and these differences being revealed in the vocabulary used to record the crime within the criminal records. Language Modelling does allow us to firstly establish that there are differences in behaviour between offenders of different sex and age (above and below 18). This is achieved by defining a separate language model per offender category. By exploring the differences probabilities assigned to the terms for, say, the male and female model, we have discovered the terms and thus the behavioural features which are more likely to occur in one group or the other.

References

1. Canter, D., Fritzon, K.: Differentiating arsonists: A model of firesetting actions and characteristics. Legal and Criminal Psychology 3, 73–96 (1998)
2. Bache, R., Crestani, F., Canter, D., Youngs, D.: Application of Language Models to Suspect Prioritisation and Suspect Likelihood in Serial Crimes. In: International Workshop on Computer Forensics (to appear, 2007)
3. Bai, J., Nie, J., Paradis, F.: Text Classification Using Language Models. In: Asia Information Retrieval Symposium, Poster Session, Beijing (2004)
4. Peng, F., Schuurmans, D.: Combining naive Bayes and n-gram language models for text classification. In: Twenty-Fifth European Conference on Information Retrieval Research (2003)
5. Peng, F., Schuurmans, D., Wang, S.: Augmenting Naive Bayes classifiers with statistical language models. Information Retrieval 7(3), 317–345 (2003)