A Continuous Language Modelling Approach for Assessing Real-valued Attributes of Documents

Richard Bache1 and Fabio Crestani[2]

[1]Department of Computer and Information Science,
University of Strathclyde, Scotland
richard.bache@cis.strath.ac.uk
[2]Faculty of Informatics
University of Lugano, Lugano, Switzerland
fabio.crestani@unisi.ch

Abstract: Probabilistic Language Modelling has been widely used for document classification and document ranking. This paper focuses on the classification task. Since language models are generative models, each language model is assumed to emit terms with (fixed) probabilities that can notionally generate the document to be classified. The fixed probabilities are derived from a training set of documents. Such an approach assumes a finite number of competing models representing each category or alternatively two models representing the presence or absence of some Boolean attribute. However, it has been widely recognised that a set of fixed models cannot cope adequately with either fuzzy classification or real-valued attributes. We therefore propose a continuous language model where the probability that a term is emitted is determined by a real-valued parameter. This model uses logistic regression to estimate a function of this parameter for each term in the vocabulary. Using just one continuous model and Bayesian approach is it possible to estimate the parameter value, representing the attribute in question, for a given unclassified document. Although other diverse applications are also mentioned, the research was motivated by the analysis of crime data. Therefore, in the experimentation presented in this paper, documents are descriptions of crimes and the attributes of interest are age of offender and distance travelled from the offender's home, as examples of two continuous attributes. Estimates from the model shows to correlate significantly with actual values establishing a relationship between the behaviour described in the crime report and both age and distance travelled.

1. Introduction

Police routinely collect large quantities of crime data and store this electronically. Such data consists typically of descriptions of how a crime was committed, comprising free text and structured data. Where a crime has been solved, further data is then available about the offender. Such data archives lend themselves to data mining techniques to identify relationships between crimes and the characteristics of offenders such as age, sex, and ethnic group or where they lived.

Our motivation for conducting this investigation is to develop automatable quantitative techniques for relating features of the crimes to the kinds of people who may have committed them; this is often referred to as *offender profiling*. Traditionally, criminal profiling has been conducted subjectively with little empirical basis. More recent work has sough to apply a quantitative approach such as Canter and Fritzon's analysis of arson [5] where one of the characteristics of interest was age. Given that this type of analysis of crime data is still in its infancy, finding even weak statistical relationships would be considered a useful result and could spur further investigation. Furthermore, if we used transparent techniques where inferences can be linked back to the underlying variables then this enables us to explore which features differentiate kinds of offenders. It should be noted that where such techniques were used by law enforcement agencies such as the police, transparency is essential in order to justify the conclusions. The fact that some opaque mathematical model has made some inference would not convince senior officers or courts of law.

The work presented here was carried out in the context of the project, iMOV (Interactive Modus Operandi Visualisation), funded by EPSRC (the UK Engineering and Physical Sciences Research Council). The project is a joint work of the Department of Computer and Information Sciences of the University of Strathclyde, Glasgow, and the Department of Investigative Psychology of the University of Liverpool. This particular study wished to investigate the relationship between the behaviour of the offender when committing the crime and both the age and distance travelled from the home base. If we consider the description of the offender's behaviour (free text and other variables) to comprise a (virtual) document describing the crime, then we can see that relating the behaviour to a characteristic such as gender or racial group of the offender is a document classification problem. However, since age and distance travelled are not naturally categorical data, attempts to 'classify' on these bases would require defining arbitrary categories e.g. *young* and *old* or *near* and *far* with an arbitrary age or distance boundary.

Bayesian approaches have been used previously to analyse crime data [11]. Language Modelling is a Bayesian approach and has been applied widely to the problem of document classification [4, 12, 13]. Specifically in the crime domain, this approach has been used to link the behaviour of known offenders [1] and also for identifying Boolean characteristics of offenders [2]. The fact that probabilities are associated with each term means that it is possible to peek inside the model and relate terms in the document (e.g. words or behavioural features) to characteristics of the offender. Language models in this respect offer advantages over other classification techniques such as vector support machines or neural nets which are essentially opaque. The drawback of the traditional Language Modelling approach is that it assumes that a document is placed in exactly one of a finite set of collectively exhaustive and mutually exclusive categories. Since distance and age are defined on a numeric scale, existing language models cannot deal adequately with a continuous attribute.

We will henceforth refer to traditional language models as *constant* models. Although they represent a stochastic process, the probabilities associated with each term in the vocabulary are assumed to be fixed for each model, having been first calibrated on some training set. Although it would be possible to apply two or more constant models over intervals of a numeric scale, such an approach has three drawbacks:

- The output can only be a range of values corresponding to a discrete category rather than a point estimate or a confidence interval.
- The model assumes catastrophic changes at (often arbitrary) boundaries and so training points at these boundaries will contribute 100% to either one language model or the other.
- Any attempt to make the intervals smaller necessarily requires increasing the number of models and spreading the training data more thinly.

Here we propose an alternative approach. We assume a single language model that has continuous parameters. These parameters alter the probabilities of each term in the vocabulary since each term has associated with it a probability function rather than a fixed probability. Using a Bayesian approach we then seek to estimate the value of these parameters for a document for which the attribute is unknown. We called this class of models: *continuous language models.* We will consider here the specific case where there is a single parameter representing the attribute of interest although such a method could, in principle, be extended to multiple parameters to deal with situations where many attributes are being estimated simultaneously.

The models were presented by Bache and Crestani in [3]. This paper extends [3] by explaining the models in more detail and by presenting an extensive evaluation of their effectiveness. The potential of these models extends well beyond the analysis of crime data. There will be situations where some attribute of a text is not dichotomous (e.g. reading age or degree of sexual or violent content) or where classification may be fuzzy (e.g. news stories that are in varying degrees about two topics). However, these applications are not considered here.

The rest of the paper is organised as follows. Section 2 describes the continuous language model by analogy with classification by discrete language models. In section 3 we describe the crime data used for the analysis. Section 4 presents the empirical results by comparing the continuous models with the alternative of using a dichotomous set-up with an arbitrary cut-off point. In Section 5, we show how the extraction of tokens can show differences in behaviour between older and younger offenders as well as those who travel near or far. Section 6 offers some conclusions.

2. Deriving Continuous Language Model

We can identify three distinct steps when applying traditional Language Modelling to a classification problem and this provides a useful template for explaining continuous language models.

- The training data is used to calibrate one language model per category, that is to assign probabilities for each term in each model. Smoothing is usually applied at this stage to account for the terms that do not appear in a specific document (see later).
- For a document D, we calculate the probability $P(D \mid M_i)$ that each of the models M_i could have generated that document. This is achieved by multiplying the probabilities for each term in the vocabulary. Here *term* refers to words or phrases from free text or the name of a category for categorical data.

- Bayes' Theorem is then used to invert the probabilities by assuming some prior, so that we can calculate a probability that a particular model generated a given document, that is $P(M_i \mid D)$.

Continuous language models seek to estimate some attribute relating to the document and thus we assume instead a single language model with a parameter v representing age, distance or some other quantity. Without loss of generality, we will henceforth consider age as the parameter of interest. Using the above template, firstly, we use logistic regression to assign a function to each term in the vocabulary; this function yields a probability for a given age. The second step produces a probability of the document being generated as a function of age by multiplying the probabilities of each term. Finally we use the continuous version of Bayes' Theorem by assuming a prior distribution of ages over solved offences and calculating the posterior distribution of the age of the offender of the crime under analysis. A point estimate can be made by, for example, taking the mean of this distribution.

2.1 Multinomial versus Multiple Bernoulli

It is worth noting here that, whereas it is usual to speak of a language model generating a document, strictly speaking the model actually generates the document's index. Stopwords are removed, words may be stemmed or lemmatised and word order is lost to yield a bag of terms; categorical variables are added as extra words or phrases to the text. It is because we are using a bag of terms that the language models used are unigram and thus take no account of context or word order.

We argue that applications of Language Modelling to document classification and to document retrieval are closely related. We can thus divide the traditional language models used in both into two groups: multinomial and multiple Bernoulli. The former takes into account the number of incidences of a term in the index and the latter considers only whether a term is present or absent. The seminal paper by Ponte and Croft [14] used Language Modelling for document retrieval and proposed the Bernoulli approach although many recent developments tended to favour multinomial models [5]. Losada [9] argues that in sentence retrieval, Bernoulli models can offer an advantage. McCallum and Nigam [10] show that Bernoulli models works better where the vocabulary is smaller. It is worth noting that for the data considered here the vocabulary is small and, in common with sentence retrieval, the documents are short (typically 20 words). It is also true that words (other than stopwords) rarely appear more than once in a document. All of these points would gravitate towards using a Bernoulli approach. However, Bernoulli models also offer a further advantage in that the probability of each term being emitted is independent of all probability of other terms. This is not true of the multinomial models and would make adaptation to a continuous model more problematic. We now explain this point in more detail.

For all unigram language models, the emission of any term is assumed to be independent of any other. However multinomial models assume that the document comprises *n* terms emitted randomly from the vocabulary set with replacement. Thus a probability $P(t_i)$ is assigned to each term in the vocabulary V. It follows that:

$$\sum_{t_i \in V} P(t_i) = 1 \tag{1}$$

Therefore, for any one term to become more frequent, at least one term other must become less frequent. Multiple Bernoulli models assume independent trials in which each specific term in the vocabulary is either emitted or not. Thus one term can become more frequent while not affecting the frequency of any other terms. In other words, it is possible to estimate the frequency of one term independently of all the others. It is this property of the Bernoulli approach that allows us to construct a continuous model.

2.2 Calibrating the Model

Since each term can be treated separately, we can express the probability of each term in the vocabulary as a function of the parameter v representing the age of the offender:

$$P(t_i) = f_i(v) \tag{2}$$

This function, which yields a value between 0 and 1, must be estimated from some set of training data. For each document in that set either a term will be present or absent. So, since we wish to estimate a probability function from Boolean data, we use logistic regression with an affine function:

$$\log it(P(t_i)) = \log\left(\frac{P(t_i)}{1 - P(t_i)}\right) = a_i v + b_i \tag{3}$$

It is then possible to estimate the values of a_i and b_i so that for a given v we can calculate the probability of term t_i appearing in the document. However, there are two problems that arise from the use of logistic regression because it necessarily requires a numerical approximation algorithm.

Firstly, there may be terms which are either always present in or always absent from every training document. Any logistic regression algorithm will fail to estimate a_i and b_i in this situation. The maximum likelihood estimate of the probabilities of these terms will be 1 and 0 respectively for any value of v. However, this will lead to a problem known to affect traditional language models, the *Zero-Probability Problem* (ZPP). If a language model gives a probability of zero to a term then any document which contains that term cannot, by definition, be generated by the model. Thus the probability of generation will be zero. Similarly if the probability is 1 then any document that fails to contain it also cannot be generated. The usual solution to this problem is smoothing but here we shall show that smoothing is unnecessary. Clearly, if a term is always present or absent in the training documents then we have to assume it is independent of v. We can then assign some constant probability to that term, such as:

$$P(t_i) = k_i \tag{4}$$

However, it is easy to show that any constant term it would cancel out in the numerator and denominator of Bayes' theorem. Since the value of the constant assigned is irrelevant, we can simply disregard the term altogether.

The second problem occurs for a small minority of sparse terms where the logistic regression algorithm fails to converge. This will typically be no more than two or three terms in a vocabulary of thousands. Inspection indicated that these are terms that occurred in only one document, were unlikely to occur in the test set and had little behavioural importance anyhow. Thus the pragmatic solution was to remove these terms from the analysis too. An investigation into various logistic regression algorithms may reduce this, although there is no guarantee that the issue can be eliminated with certainty. Thus the elimination of a small minority of problematic terms may always be necessary.

For terms, for which a_i and b_i could be estimated, $f_i(v)$ will always produce a value such that $0 < f_i(v) < 1$ for any value of v. Thus ZPP cannot occur with this model. Nevertheless, it could be argued that smoothing could improve the performance of the model. However, when applied empirically, the smoothing did not show any improvement in performance.

2.3 Calculating the Probability of the Attribute

The probability of the language model generating the document (or, strictly speaking, its index) for a given value of v may be written as:

$$P(D \mid v) = \prod_{t_i \in D} P(t_i) \times \prod_{t_i \notin D} (1 - P(t_i)) \tag{5}$$

Since v is modelled as continuous, the continuous for of Bayes' theorem gives us:

$$pdf(v \mid D) = \frac{pdf(v) \cdot P(D \mid v)}{P(D)} \tag{6}$$

Here $P(D)$ can be calculated by assuming that the area under the distribution $pdf(v \mid D)$ will be 1. The prior distribution $pdf(v)$ can be estimated from the training data. Since here we wish only to calculate the mean of the posterior distribution, the fact that this consists of a number of discrete points and is thus rather lumpy does not matter. However, if we wished to produce a graphical representation of the likely ages (or distances), a smoothed curve could be fitted.

3. Data Analysed

The data was drawn from a digital police archive from a large city for crimes reported over a period of years. Nine datasets of various crime types were available with age

data and hailed from one district of the city over a four-year period. One set of burglary data contained coordinates of both crime scene and the offender's home base and thus Euclidean distance could be calculated. It covered the whole city but this was restricted to prolific burglars over more than 10 years. These constraints were imposed by the availability of the data and were outside the control of the researchers. Only solved crimes were considered for the analysis. For the age data, crimes with more than one offender were removed since there would be more than one age associated with that crime. For the distance data there was always one offender per crime. Table 1 summarises the data sets and their respective sizes.

Table 1: Summary of Datasets.

Set No.	Crime Type	Continuous Variable	Median Value	Num Points	No. Offenders
1	Theft from Vehicles	age	16	248	159
2	Other Theft	age	25	326	284
3	Shoplifting	age	29	2060	1618
4	Assault	age	31	2073	1881
5	Criminal Damage	age	26	849	724
6	Damage to Vehicles	age	27	220	207
7	Burglary	age	27	1126	556
8	Street Robbery	age	18	263	210
9	Sexual offences against women	age	37	123	117
10	Burglary	distance (meters)	2689	1376	83

Words from free text were lemmatised using a lemmatiser based on WordNet [7]. Stopwords from a standard list were removed. Certain addition words which may identify age or sex (e.g. *young* or *female*) were removed. Others, which revealed either age or sex, were mapped to neutral terms (e.g. *child*, *man*, *woman* to *person*). Codes were of two types. Allegation codes indicated the type or subtype of offence and one was always present. Features codes related to some observed behaviour and any number could be present although the typical number was one or two. The codes were mapped to phrases which were hyphenated and marked with a $-sign to remain separated from free text words.

4. Empirical Validation

The purpose of the experiments presented here was to demonstrate firstly that the continuous language models have predictive power and secondly that they perform comparably with a dichotomous model comprising two constant language models splitting the continuous variable at the median. It can be argued that a model that yields a numeric value or a distribution of numeric values is intrinsically preferable than one that assigns categories so it would be sufficient to show that it performs as well as a dichotomous model. Our third purpose was to compare two varieties of dichotomous models, multinomial and multiple Bernoulli. If the multinomial model outperformed the Bernoulli model then it would call into question the Bernoulli assumption for the

continuous model. Both constant types of constant language model were applied with Jelinek-Mercer smoothing [8] taking $\lambda = 0.5$.

4.1 Experimental Design

A common strategy when testing models such as the ones presented here is to assign the data randomly into a training set and a test set, typically with 50% of the points in each. However such an approach was shown to give very different results for each random allocation. Given the relatively small size of the datasets, it was practical to use the *leave-one-out* (LOO) or jack knife strategy of selecting training and test data. Usually, this technique consists of removing one point from the data set and using all the remaining data to training the model. The model is then tested on the removed data point. This procedure is repeated for each data point until each point has been excluded precisely once. However, this potentially leads to the problem of serial crimes where several crimes are committed by the same offender and thus the training set contains crimes by the same offender as the test set. Although we expect to find commonalities in behaviour between people of the same age, this will never be as strong as the consistency of behaviour of one individual. So if a 25-year-old burglar has a data point in both the training set and the test set then the model may well identify a connection peculiar to that individual rather than to offenders of that age in general.

One possible solution to the serial crime problem was to remove the serial crimes. For many of the data sets this would have been a viable option although it would reduce the size of the datasets. For data set 7 and particularly data set 10 of Table 1 this would create a problem since it lost more than half the data points. Thus the solution was to use a *leave-one-offender-out* strategy. Here the test set comprises all the offences by one offender and the training set comprises all offences committed by other offenders. This procedure is then run once for each offender.

The two types of model yield different results. The dichotomous models predict a category whereas the continuous model predicts a distribution of values of which we may take some measure of central tendency such as the mean. Nevertheless, there should be a correlation between the predicted and actual values although different statistical tests would be applied. For the dichotomous model we use the Chi-squared test. For the continuous model we use Pearson's correlation coefficient between the mean estimated age and the actual age.

4.2 Results

Table 2 shows the significance of correlation for both the dichotomous models and the continuous model using one-sided tests. The fact that the models detected significances also demonstrates that age is a factor in the behaviour of the offender in all but one dataset. Dataset 9 is particularly small and this may explain why no significant correlation was found. The continuous model shows significance in all other cases and therefore outperforms either of the two other models. These models fail to find a significant relationship in burglary either based on age or distance travelled. From these

data we can conclude that the continuous model is capable of finding relationships that a dichotomous model cannot.

Table 2: Significance of correlation (better than 5% shown in bold).

Set No.	Crime Type	Variable	Dichotomous Models – Chi squared		Continuous Model --Pearson
			Multivariate	Multiple Bernoulli	
1	Theft from Vehicles	age	**0.015**	**0.002**	**0**
2	Other Theft	age	**0.046**	**0.022**	**0.013**
3	Shoplifting	age	**0**	**0**	**0**
4	Assault	age	**0**	**0**	**0**
5	Criminal Damage	age	**0**	**0**	**0.026**
6	Damage to Vehicles	age	**0.006**	**0.002**	**0.001**
7	Burglary	age	0.155	0.081	**0.007**
8	Street Robbery	age	**0**	**0**	**0**
9	Sexual offences against women	age	0.266	0.399	0.353
10	Burglary	distance	0.4	0.537	**0**

5 Exploring the Contribution of Individual Terms

As mentioned in Section 1, the model proposed here has a degree of transparency so that it is possible to determine what impact each term in the vocabulary has on implying either an older or younger offender. This has two possible uses:

- For any unsolved crime, we can determine whether each term in the document (excluding stopwords) implies younger or older offenders thus shedding some light as to why the model has estimated a given age. It would also indicate whether the terms were concordant in this implication or were giving mixed messages.
- By looking at the offences together, we can indicate which terms tend to imply older or younger behaviour in general. This information has a qualitative application for law enforcers to apply results of the analysis in the field and well as being interest to investigative psychologists.

The first use could be achieved by displaying a crime report with the various terms colour coded depending on the extent that term relates to younger or older behaviour, as shown in Figure 1.

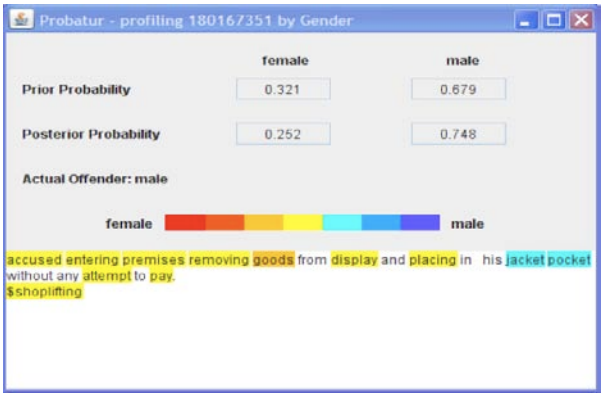

Figure 1: Colour coding of police report to indicate evidence of the sex of the offender.

Here we consider a second use in more detail although both uses require a measure of age sensitivity defined as follows. Given that each term has a probability that is a function of age, we calculate the derivative of the probability with respect to age. A strongly negative value indicates it is more common amongst younger offenders and a strongly positive value will indicate it relates older behaviour. Values around zero suggest that such a feature is not influenced much by age. Where we have terms from free text, then the relationship between identifiable features and words used is a complex one because of synonymy, polysemy and words that do not relate to behaviour at all such as proper nouns. Nevertheless, ranking the vocabulary by this derivative does yield interesting and intuitive results. Rearranging equation 3 gives:

$$P(t_i) = \frac{e^{av+b}}{1+e^{av+b}} \tag{7}$$

and differentiating it with respect to *v* yields:

$$\frac{dP(t_i)}{dv} = \frac{ae^{av+b}}{\left(1+e^{av+b}\right)^2} \tag{8}$$

There is the problem of serial offences when performing this analysis. A particular term may relate to the behaviour of a single offender. A prolific offender with a unique recurring feature may lead us erroneously to infer that it was common to all offenders of that age. Thus serial crimes were removed so that there was exactly one crime per offender.

Table 3 shows the top and bottom ranked terms for assault. Note that terms starting with a $-sign are derived from codes and also that *parent*, *spouse* and *offspring* are gender-neutral terms for *mother*, *father* etc. Inspection of the terms shows that whereas young offenders are involved in more serious assaults (e.g. Section 18 assault which can carry life imprisonment) with punching and kicking and attacks involving the head being common features, older offenders are more likely to be involved in domestic incidents resulting in less serious assaults on family members.

Analysis of other data sets reveals interesting patterns too. Younger offenders are more likely to burgle non-residential premises whereas older burglars target homes. Burglars who operate close to their home base are more likely to climb into properties or enter through doors; those who have travelled further tend to enter through windows and then conduct an untidy search. For older offenders, damage to vehicles is more likely to occur as a result of a road rage incident whereas younger offenders are more likely to vandalise an unattended vehicle. A similar pattern emerges for criminal damage where older offenders damage property of an individual with whom they have a dispute. Younger offenders are more likely to engage in 'victimless' acts of vandalism (i.e. not directed at an individual) such as spray painting public property. In shoplifting, theft of bottles of spirits (whisky and vodka) is associated with older offenders.

Table 3 – Terms Ranked by Derivative of Probability with regards to Age.

Rank	Lowest Ranked (young)		Highest Ranked (old)	
	Word	Value	Word	Value
1	$victim-kicked	-0.00323	$common-assault	0.00203
2	kick	-0.00236	argument	0.00203
3	victim	-0.00186	parent	0.00144
4	$victim-punched	-0.00171	spouse	0.00130
5	punch	-0.00166	slap	0.00113
6	suspect	-0.00138	offspring	0.00106
7	head	-0.00114	grab	0.00105
8	attack	-0.00110	arm	0.00092
9	$assault-section-eighteen	-0.00103	domestic	0.00077
10	$sharp-instrument	-0.00085	$victim-threatened	0.00075

These results were validated by police officers and considered extremely interesting by the investigative psychologists that were our partners in the iMOV project. They are currently subject to further analysis and validation by field studies.

6 Conclusions

The experimental results show that the continuous language model is able to produce estimates of age and distance for a notionally unsolved crime that correlate significantly with the actual quantities. This model outperforms the alternative dichotomous model. The use of multiple Bernoulli models is appropriate for the nature of the data analysed since the dichotomous Bernoulli model performs as well as the multinomial one. The fact that the models exhibit a degree of transparency is shown to be useful in both explaining an inference of the model but also to identify the different styles of behaviour related to different ages or distances travelled.

Acknowledgements

We would like to thank Prof. David Canter and Dr. Donna Youngs of the Department of Investigative Psychology of the University of Liverpool for many interesting discussions, which inspired the development of the models presented in this paper, and also for providing the data on which the models were tested.

References

1. Bache, R., Crestani F., Canter D., Youngs D., Application of Language Models to Suspect Prioritisation and Suspect Likelihood in Serial Crimes, *International Workshop on Computer Forensics*, pages 399-404, Manchester, UK, 2007.
2. Bache, R., Crestani F., Canter D., Youngs D., Mining Police Digital Archives to Link Criminal Styles with Offender Characteristics, *International Conference on Asian Data Libraries (ICADL),* pages 493-494, Hanoi, Vietnam, 2007.
3. Bache, R., Crestani F., Estimating Real-valued Characteristics of Criminal from their Recorded Crimes. *The Seventeen ACM Conference on Information and Knowledge Management (CIKM)*, pages 1385-1386, Napa Valley, USA, 2008.
4. Bai J., Nie J., Paradis F., Text Classification Using Language Models. *Asian Information Retrieval Symposium (AIRS)*, Poster Session, Beijing, 2004.
5. Canter, D., Fritzon, K., Differentiating arsonists: A model of firesetting actions and characteristics, *Legal and Criminal Psychology*, vol. 3, pp 73-96, 1998
6. Croft, W.B., Lafferty J, *Language Modeling for Information Retrieval*, Kluwer Academic, 2003.
7. Fellbaum C. (Ed.): WordNet – An Electronic Lexical Database, MIT Press, 1998.
8. Jelinek, F., Mercer, R., Interpolation estimation of Markov source parameters from sparse data. *Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands, 1980.
9. Losada D., Language Modeling for Sentence Retrieval: A comparison between Multiple-Bernoulli Models and Multinomial Models, *Information Retrieval Workshop*, Glasgow, Scotland, 2005.
10. McCallum A., Nigam, K., A Comparison of Event Models for naïve Bayes Text Classification, *Proc. AAAI/ICML-98 Workshop on Learning for Text Categorisation*, pages 41-48, Madison, Wisconsin, 1998.
11. Oatley G.G., Ewart B., Crimes Analysis Software: Pins in Maps, Clustering and Bayes Net Prediction. *Expert Systems with Applications*, 25(4):569-588, 2003
12. Peng, F., Schuurmans, D., Combining naive Bayes and n-gram language models for text classification, in *Twenty-Fifth European Conference on Information Retrieval Research (ECIR),* peges 335-350, Pisa, Italy, 2003.
13. Peng, F., Schuurmans, D.,Wang, S. Augmenting Naive Bayes classifiers with statistical language models. In *Information Retrieval,* 7(3):317-345, 2003.
14. Ponte J.M., Croft W.B., A Language Modeling Approach to Information Retrieval, in *Proceedings of the Twenty First Conference on Information Retrieval Research (SIGIR)*, Melbourne, Australia, page 275-281, 1988.