Information Processing and Management 49 (2013) 341–354

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Authorship attribution based on a probabilistic topic model

Jacques Savoy

Computer Science Department, University of Neuchatel, Rue Emile Argand 11, 2000 Neuchâtel, Switzerland

ARTICLE INFO

Article history:
Received 3 February 2012
Received in revised form 4 May 2012
Accepted 25 June 2012
Available online 31 July 2012

Keywords:
Authorship attribution
Text categorization
Machine learning
Lexical statistics

ABSTRACT

This paper describes, evaluates and compares the use of *Latent Dirichlet allocation* (LDA) as an approach to authorship attribution. Based on this generative probabilistic topic model, we can model each document as a mixture of topic distributions with each topic specifying a distribution over words. Based on author profiles (aggregation of all texts written by the same writer) we suggest computing the distance with a disputed text to determine its possible writer. This distance is based on the difference between the two topic distributions. To evaluate different attribution schemes, we carried out an experiment based on 5408 newspaper articles (*Glasgow Herald*) written by 20 distinct authors. To complement this experiment, we used 4326 articles extracted from the Italian newspaper *La Stampa* and written by 20 journalists. This research demonstrates that the LDA-based classification scheme tends to outperform the Delta rule, and the χ^2 distance, two classical approaches in authorship attribution based on a restricted number of terms. Compared to the Kullback–Leibler divergence, the LDA-based scheme can provide better effectiveness when considering a larger number of terms.

1. Introduction

In order to manage the huge amount of freely available textual information, various text categorization tasks have been proposed. In this study, we address the authorship attribution (AA) problem (Love, 2002; Juola, 2006) whereby the author of a given text must be determined based on text samples written by known authors. Knowing that the real author is one of the candidates, this specific challenge is defined as the *closed-class authorship attribution* problem. In such applications, the query text might correspond to various items such as a romance, a part of a play, an anonymous letter, a web page, or a sequence of paragraphs. Other questions are related to such issues as the mining of demographic or psychological information on an author (*profiling*) (Argamon et al., 2006) or simply determining whether or not a given author did in fact write a given Internet message (chat, e-mail, Wikipedia article) or document (*verification*) (Koppel, Schler, & Argamon, 2009).

As with all text categorization problems, the first step is to represent the texts by means of numerical vectors comprising selected features that help in discriminating among the various authors or categories. In the current context, we must identify pertinent terms depicting differences between the authors' writing styles. In the second stage, we weight the chosen features according to their discriminative power as well as their importance in the text representation. Finally, through computing a distance or applying classification rules, the system assigns the most appropriate author to a given input text (*single-label categorization* problem).

In the classical authorship attribution studies (Juola, 2006), we usually focus on frequent words or on a small number of very frequent terms to represent each text item. We then define a distance measure and determine the probable author of the query text as the one that depicts the smallest distance. As an alternative way, the machine learning paradigm (Sebastiani, 2002) will focus more of the selection process on identifying the most pertinent features according to their

E-mail address: Jacques.Savoy@unine.ch

0306-4573/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2012.06.003

distribution into the different categories or authors. As classification model, this paradigm may use a larger range of possible strategies (Witten & Franck, 2005). In this paper we propose following a third way. Using a probabilistic generative approach based on a probabilistic topic model (Blei, Ng, & Jordan, 2003) we show how we can determine the possible author of a disputed text.

The rest of this paper is organized as follows: Section 2 exposes the state of the art. Section 3 describes the corpora used in our experiments while Section 4 presents an overview of four authorship attribution models used as baselines. Section 5 presents the idea of latent Dirichlet allocation (LDA) and its application as an authorship attribution method. An evaluation of these classifiers is presented and analyzed in Section 6.

2. Related work

Authorship attribution owns a long-standing history (Mosteller & Wallace, 1964, Juola, 2006, Zheng, Li, Chen, & Huang, 2006, Stamatatos, 2009). As a first solution, past studies have proposed using a unitary invariant measure that must reflect the style of a given author but should vary from one writer to another. The average word length, mean sentence length, as well as Yule's *K* measure and statistics based on type-token ratios (e.g., Herdan's *C*, Guiraud's *R* or Honoré's *H*) (Baayen, 2008) have been suggested as well as the proportion of word types occurring once or twice. However, none of these measures has proved satisfactory (Hoover, 2003).

As a second approach, multivariate analysis (principal component analysis (Craig & Kinney, 2009), cluster analysis, or discriminant analysis (Jockers & Witten, 2010)) has been applied to capture each author's discriminative stylistic features. In this case, we represent documents as points within a given lexical space. In order to determine who might be the author of a new text excerpt we simply search the closest document assuming that the author of this nearest document would probably be the author of the disputed text.

Following the idea of measuring an intertextual distance, some recent studies suggest using more topic-independent features that may reflect an author's style more closely. In this vein, we can limit text representation to function words (e.g., determiners, prepositions, conjunctions, pronouns, and certain auxiliary verbal forms). Since the precise definition of function words is questionable, a wide variety of lists have been proposed. Burrows (2002), for example, lists the top *m* most frequent word types (with $m = 40–150$), while the list compiled by Zhao and Zobel (2005) contains 363 English words. Not all studies, however, suggest limiting the possible stylistic features to a reduced set of functional words or very frequent word types. In their study of the 85 *Federalist Papers*, for example, Jockers and Witten (2010) derive 2907 words appearing at least once in texts written by all three possible authors. As another example, Hoover (2007) suggests considering up to the top 4000 frequently occurring words, including in this case both function and lexical words (nouns, adjectives, verbs, and adverbs). On the other hand, more sophisticated intertextual distances have been suggested (Labbé, 2007), where the distance between two documents depends on both their shared vocabulary and occurrence frequencies.

As another source of features, we could take into account part-of-speech (POS) tags by measuring their distribution, frequency and patterns or their various combinations. Finally, some studies, usually related to the web, suggest exploiting structural and layout features (the number of lines per sentence or per paragraph, paragraph indentation, presence of greetings, etc.). Additional features that could be considered are particular orthographic conventions (e.g., British vs. US spelling) or the occurrence of certain spelling errors.

As a second main paradigm, we can apply machine learning techniques to determine the probable author of a disputed text. In this vein, we can see each author as one possible category using the set of previously described features. We then need to define a classification model that can distinguish among possible authors. Zheng et al. (2006) suggests employing decision trees, back-propagation neural networks and support vector machines (SVMs). They found that by solely using lexical features, the performance levels obtained are similar to those of POS and lexical feature combinations. This finding is confirmed by another recent study (Zhao & Zobel, 2007). Zheng et al. (2006) also found that SVM and neural networks tend to have significantly better performance levels than those achieved by decision trees. Nanavati, Taylor, Aiello, and Warfield (2011) have obtained good overall performance with a naïve Bayes classifier for deanonymizing referees' reports extracted from two scientific conferences. Zhao and Zobel (2005) on the other hand, found that the Nearest Neighbor (NN or *k*-NN) approach tends to produce better effectiveness levels than both the naïve Bayes and decision-tree approaches.

Instead of applying a machine learning classification method, Burrows (2002) designed a more specific Delta classifier based on the differences of standardized word occurrence frequencies. This method assumes that authors' styles are best reflected by identifying the use of function words (or very frequent words) rather than relying on a single vocabulary measure or more topic-oriented terms. Recently, Jockers and Witten (2010) showed that the Delta method could surpass performance levels achieved when using the SVM method.

3. Evaluation corpora

The number of publicly available test corpora related to the authorship attribution domain is quite limited. Thus, making sufficiently precise comparisons between reported performances and general trends regarding the relative merits of various classification approaches is problematic. The relatively small size of the corpora used is a second concern. In various experiments, the number of disputed texts and possible authors are rather limited (e.g., the *Federalist Papers* are composed of 85 texts from which 12 are disputed articles mainly written by two possible authors (Mosteller & Wallace, 1964, Jockers & Witten, 2010)).