Information Processing and Management 53 (2017) 1103–1119

Contents lists available at ScienceDirect

Information Processing and Management

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

Distance measures in author profiling

Mirco Kocher, Jacques Savoy*

University of Neuchatel, rue Emile Argand 11, 2000 Neuchatel, Switzerland

ARTICLE INFO

Article history:
Received 29 November 2016
Revised 4 April 2017
Accepted 19 April 2017

Keywords:
Distance measure
Author profiling
PAN-CLEF
Text categorization

ABSTRACT

Determining some demographics about the author of a document (e.g., gender, age) has attracted many studies during the last decade. To solve this author profiling task, various classification models have been proposed based on stylistic features (e.g., function word frequencies, *n*-gram of letters or words, POS distributions), as well as various vocabulary richness or overall stylistic measures. To determine the targeted category, different distance measures have been suggested without one approach clearly dominating all others. In this paper, 24 distance measures are studied, extracted from five general families of functions. Moreover, six theoretical properties are presented and we show that the Tanimoto or Matusita distance measures respect all proposed properties. To complement this analysis, 13 test collections extracted from the last CLEF evaluation campaigns are employed to evaluate empirically the effectiveness of these distance measures. This test set covers four languages (English, Spanish, Dutch, and Italian), four text genres (blogs, tweets, reviews, and social media) with respect to two genders and between four to five age groups. The empirical evaluations indicate that the Canberra or Clark distance measures tend to produce better effectiveness than the rest, at least in the context of an author profiling task. Moreover, our experiments indicate that having a training set closely related to the test set (e.g., the same collection) has a clear impact on the overall performance. The gender accuracy rate is decreased by 7% (19% for the age) when using the same text genre during the training compared to using the same collection (leaving-one-out methodology). Employing a different text genre in the training and in the test phases tends to hurt the overall performance, showing a decrease of the final accuracy rate of around 11% for the gender classification to 26% for the age.

1. Introduction

In our digital world, author profiling and authorship attribution are viewed as important questions from a security perspective or regarding the increased number of pseudonymous posts and messages (Olsson, 2008). In literary studies, being able to verify the gender of a given character may open new research directions (e.g., is Juliet really a female figure? (Craig & Kinney, 2009)).

To solve these questions, various approaches have been suggested based on vocabulary richness measures (Holmes, 1998), (Baayen, 2008), stylometric similarities (Burrows, 2002; Savoy, 2012), or machine learning models (Stamatatos, 2009; Jockers & Witten, 2010). In many cases, texts are represented by vectors in which the different dimensions correspond to words, characters, *n*-grams of letters or words, part-of-speech (POS) categories, or other possible stylistic measures (e.g., sentence

* Corresponding author.
E-mail addresses: Mirco.Kocher@unine.ch (M. Kocher), Jacques.Savoy@unine.ch (J. Savoy).

http://dx.doi.org/10.1016/j.ipm.2017.04.004
0306-4573/© 2017 Elsevier Ltd. All rights reserved.

length, lexical density, etc.). These models assume usually that the corresponding dimensions are orthogonal and the number of dimensions varies widely from one model to another (e.g., 3 in (Fung, 2003), 10 in (Mosteller & Wallace, 1964), 2907 in (Jockers & Witten, 2010), and more than 73,000 in (Khonji & Iraqi, 2014)).

To define the exact demographic category of the author, several proposed approaches need to compute a distance (or similarity) measure between the query text and the representations of the different categories. The shortest distance (or the maximum similarity) determines the predicted class. The choice of the distance measure is often based on *ad hoc* considerations, tradition, or limited empirical evidence.

The objectives of this paper are the following three. First, we want to establish a set of useful properties that a distance measure must respect. Second, and based on a large number of different test collections, we want to determine a reduced set of distance measures showing the most effective performance. Third, using a relatively large number of test collections, we have the opportunity to quantify the influence of the training set on the test set. Thus, we want to estimate the possible performance variations when using the same collection during the training and test phases, when using different collections with the same text genre, or when there are different text genres in both stages.

The rest of this paper is organized as follows. The next section presents the state of the art in author profiling with the focus on the gender and age determination. The third section explains the distance measures and the properties we can expect from an effective one in the context of authorship attribution or profiling. In the fourth section, we perform a theoretical assessment of the different distance measures. The fifth section describes the test collections and the evaluation methodology used in the experiments. The evaluation of the different distance measures is exposed in the sixth section, together with the evaluation of different combinations during the training and test phase. A conclusion draws the main findings of this study.

2. Related work

The main objective of an author profiling task is to determine, as accurately as possible, some author's demographics from text (e.g., gender, age, some personality traits, social class, native language, etc. (Argamon, Koppel, Pennebaker, & Schler, 2009)). The gender distinction might be viewed as the simplest one. The classification decision can be binary and a relatively large amount of data can be collected. However, such a classification system can be effective only if the writing style between genders does differ (Eckert & McConnell-Ginet, 2013) and if such stylistic differences can be detected.

Past studies tend to demonstrate that such differences do occur when considering pervasive and frequent features such as determiners, pronouns, or part-of-speech (POS) distributions. According to Pennebaker (2011), women tend to employ more personal pronouns (especially more *I* and *we*) than men (in relative frequencies, 14.2% vs. 12.7% in blog posts). The signal does not seem to be really strong, but it exists. Looking at other lexical groups, Pennebaker (2011) indicates that men tend to employ more big words (composed of six letters or more), determiners, prepositions, nouns, numbers, and swear words. On the other hand, women use more verbs, negations (e.g., never, not), cognitive words (e.g., consider, explain, think), social words (e.g., family, folks), emotion words (e.g., fears, crying, losses) (Talbot, 2010; Rangel & Rosso, 2016), and certainty words (e.g., always, must). Of course, each individual can depict a more or less strong masculine or feminine figure.

As another way to detect the author gender, Alowibdi, Buy, and Yu (2013) suggest taking account of the first names and user names both transformed into phonemes (with the set of possible phonemes limited to 40). With other languages than English, the gender detection can be determined by considering a few words (e.g., in Portuguese, thank you is *obrigado* for a man, and *obrigada* for a woman) (Ciot, Sonderegger, & Ruths, 2013).

For most of those features, simple lists of words can be created mainly because some grammatical categories such as determiners or pronouns form a closed set. Within a given language, a new preposition cannot be created. For other POS such as nouns or verbs, new instances can occur (e.g., to google). Their identification requires however a language-dependant POS tagger. As an alternative, LIWC (Linguistic Inquiry and Word Count) (Tausczik & Pennebaker, 2010) proposes a set of word lists to measure some stylistic features (e.g., determiners, personal pronouns, modal verb forms) as well as other semantic-based categories such as positive emotions or social words.

A simple count based on a single feature cannot provide a reliable measure. The text register has an impact on those predictors, as for example, pronouns are in general less frequent in a formal context. On the other hand, political speeches delivered by US presidents contain more pronouns, even when the context is official (Savoy, 2016). Therefore, generalization based on a single experiment or using a unique text register should be viewed with caution.

As expected, some topical words are used more frequently by one of the genders (e.g., sports, job, money vs. family, shopping, friends) (Schler, Koppel, Argamon, & Pennebaker, 2006). The two genders have their preferred subjects and this aspect is reflected in their lexical choice. Based on around 100,000 blog posts (50% were written by men, 50% by women), the computer can correctly classify 72% of them based on very frequent words (Argamon, Dhawle, Koppel, & Pennebaker, 2005; Argamon et al., 2009) (19,320 authors; mean text length: 7250 words). Including also the topical terms, the machine can reach an accuracy rate of 76%. In this case, men use more terms related to technology (e.g., game, software, Linux) while women prefer writing about friends and social relations (e.g., love, cute, mom). Those examples are however related to the weblog in which other lexical features can be used to discriminate between the two genders (e.g., emoticons (Crystal, 2006)). Changing the text source requires that the most discriminative topical words between the two genders should be redefined (e.g., selecting the 1000 words depicting the highest information gain ratio (Argamon et al., 2009)).