Knowledge-Based Systems 89 (2015) 134–147

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Discriminative subprofile-specific representations for author profiling in social media

A. Pastor López-Monroy [a,*], Manuel Montes-y-Gómez [a], Hugo Jair Escalante [a], Luis Villaseñor-Pineda [a], Efstathios Stamatatos [b]

[a] *Laboratory of Language Technologies, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Luis Enrique Erro No. 1, Sta. Ma. Tonantzintla, C.P. 72840 Puebla, Mexico*
[b] *Dept. of Information and Communication Systems Engineering, University of the Aegean, Karlovassi, Samos 83200, Greece*

ARTICLE INFO

Article history:
Received 8 December 2014
Received in revised form 21 March 2015
Accepted 13 June 2015
Available online 2 July 2015

Keywords:
Author profiling
Web mining
Text classification
Social media

ABSTRACT

The Author Profiling (AP) task aims to reveal as much as possible information from a given author's document (e.g., age, gender, etc.). AP is crucial for several applications, ranging from customized advertising to computer forensics, psychology, and entertainment. Nonetheless, the AP task is far from being solved, particularly in social media domains, where the nature of documents hinder the applicability of state-of-the-art text mining tools (e.g., because of spelling-grammar errors, huge vocabularies, and the presence of many out-of-vocabulary terms). Currently, most of the work in AP for social media has been devoted to the development of descriptive features, which are used under standard representations, such as the Bag-of-Words (BoW). Nevertheless, BoW-like representations have some well known shortcomings, namely: (i) the sparsity and high dimensionality of the representation, and (ii) the failure to capture relationships, other than mere occurrence, among terms. This paper focuses on the study of alternative document representations that can deal with such issues. We propose a representation for documents that capture discriminative and subprofile-specific information of terms. Under the proposed representation, terms are represented in a vector space that captures discriminative information. Then, term representations are aggregated to represent the content of a document. In this manner, documents are represented in a low-dimensional (and discriminative) space which is non-sparse. We evaluate the effectiveness of the proposed representation on several corpora from the social media domain. The proposed representation is compared to the standard BoW representation and a wide variety of state-of-the-art AP approaches. Experimental results reveal that the proposed representation outperforms most of the reference methodologies. Furthermore, we show that the proposed representation is in agreement with previous studies on handcrafted attributes for AP.

1. Introduction

Nowadays, the huge amount of information available in the Web is constantly growing. Much of this information is in plain text written by users under different contexts, for example in: social networks, forums, blogs, emails, etc. The availability of all of this information pose a challenge to researchers and practitioners on information sciences, who have to develop automated tools for the access, organization and analysis of such information.

In this regard, AP is a task that aims at analyzing text in order to obtain as much information as possible from authors [18]. For instance, personality, native language, cultural background, demographics, etc. AP has wide applicability and can have a broad impact in our lives. For example, in marketing, specific characteristics of users could be automatically detected (e.g., gender) with the aim of showing specific advertising. Similarly in business intelligence, large corporations can perform analysis in their forums and blogs in order to know which *kind* of people are interested in their products (e.g., young people). In criminal law, knowing the linguistic profile of authors of harassing messages could be valuable for finding or condemning suspects. In recent years all these applications have gained interest for the scientific community, who has dubbed these tasks as AP. This paper focuses on the AP task in the context of social media documents.

[*] Corresponding author.
E-mail addresses: pastor@inaoep.mx (A.P. López-Monroy), mmontesg@inaoep.mx (M. Montes-y-Gómez), hugojair@inaoep.mx (H.J. Escalante), villasen@inaoep.mx (L. Villaseñor-Pineda), stamatatos@aegean.gr (E. Stamatatos).

http://dx.doi.org/10.1016/j.knosys.2015.06.024
0950-7051/© 2015 Elsevier B.V. All rights reserved.

1.1. Author profiling

According to the literature, two main problems in AP have been gained interest recently: recognition of age and gender [6,25,18,38,28]. In this context, the AP task can be approached as a single-labeled classification problem, where the different profiles (e.g., *males* vs. *females*, or *teenager* vs. *young* vs. *old*) stand for the target classes. Nevertheless, this task should not be addressed exactly as other document classification tasks such as thematic classification [39] or Authorship Attribution (AA) [41]. For example, while in AA we need to model the specific writing style of *each author* by focusing in author's specific features, in AP we need to model more general sociolinguistic features that apply to *groups of authors* and indicate how they use words given their native language, genre, age, etc. [1].

Although the AP problem has been approached by many researchers, see e.g., [6,25,18,38,28], most of the work in AP has been devoted to the analysis of textual features used to represent documents. In this aspect we can differentiate two main branches on AP: classical and social media ones. Classical AP focuses on formally written documents such as books and newspaper articles, while the social media AP focuses on informal documents (e.g., blogs, forums, tweets, etc.). For the classical AP, authors have proposed interesting combinations of textual attributes ranging from lexical features (e.g., content words [1] and function words [18]) to syntactical features (e.g., POS-based features [4] and probabilistic context-free grammars [37]). However, the use of some of these features is impractical when analyzing social media. For example, it is very hard to accurately extract syntactic information from informal documents. Therefore, some works on AP for social media have found that the most useful attributes are combinations of content and stylistic features (e.g., function words) [38,15,33].

Notwithstanding the success of content and stylistic features, in AP it is the norm to use such features with standard representations from text mining. In fact, to the best of our knowledge, there is little research devoted to developing suitable representations for AP in general. Currently, the Bag of Words (BoW) is the most used representation for AP in the social media: documents are represented by the frequency of occurrence of content/stylistic terms in documents [15,33,32]. Whereas BoW has appealing properties there are two well known issues of this representation that could compromise its applicability in AP in socialmedia:

1. *High dimensionality and sparseness:* documents usually are represented in a large vector space of length equal to vocabulary size, which can hinder the use of some learning methods [16] and impact the runtime performance [17]. Also, documents usually contain a small subset of terms, resulting in sparse representations that make difficult to interpret and build accurate models for some text analysis tasks in socialmedia [31,40].
2. *Assumption of independence among terms:* No relationship, other than mere occurrence, among terms is captured by the representation. Besides, the representation does not capture associations between terms and the different profiles. Such information could be very helpful for the AP process.

Both issues become even more severe when we consider some of the most common situations when dealing with content from the social media domain: (i) highly imbalanced categories; (ii) extremely varied size documents (e.g., some with several pages, some with a few words), and (iii) a huge number of non-dictionary terms (due to the ease to write messages). These problems are in turn combined with one of the properties that make AP for social media very challenging: Internet is by nature open to everyone. Therefore, one can expect any type of content, writing style, length, and even language, mainly because each profile is formed by groups of a wide variety of authors. This inherent diversity of social media causes a *high heterogeneity* among the users belonging to the same profile. Intuitively, we can say that, even when we could have large groups of users belonging to the same profile (e.g., females), given the *high heterogeneity*, there are in fact subgroups of females that are different among themselves. For example, while the largest group of females writes about family and friends, there are other small groups of females interested in stereotypically male topics such as video games and sports. In this context, it is desirable to have a representation that capture these details of the high diversity comprising the subgroups particularities that make each profile unique. Unfortunately, in most of the previous works on AP it is assumed that there exists certain homogeneity among all documents/authors that belong to a same profile. While the assumption of homogeneity is somewhat necessary for modeling purposes,[1] intra-profile heterogeneity is the norm in the social media domain. In consequence, standard BoW representations and usual classification schemes may fail to capture discriminative information at a fine grain. Given this context, as we already mention, there is a need for a new representation that can explicitly capture information about sub-groups (i.e., sub-profiles) in order to have a finer representation of each profile; that is precisely the propose of this work.

1.2. Overview of our method

In this work we introduce a novel approach to represent documents in AP for the social media domain. The proposed representation overcomes some of the previously mentioned limitations of BoW for AP and can deal with the challenging conditions of the AP task. In our proposal, terms are first represented in a (low dimensional) vector space that accounts for the association of terms to profiles and subprofiles. This term representation aims at capturing discriminative intra-profile information that can deal with profile heterogeneity in AP. Then term representations are aggregated to characterize documents, and a standard classification procedure is adopted.

In this manner, documents are represented in a low-dimensional space which is non-sparse. What is more, by means of the aggregation procedure, document representations capture term-term relationships; whereas term-subprofile information is captured by the initial term representation. We emphasize that the latter feature of our approach can help us to deal with the variability of documents in terms of content, informality and style, as we can model term usage at a fine level. Furthermore, the proposed representation can be used with any type of features (e.g., content or stylistic based) as it lies at the top of the representation of documents; also it can be used with any classification model.

We evaluated and compared our proposal against the traditional BoW and a wide diversity of state-of-the-art approaches. For this, we performed an extensive study that comprises ten of the most important data sets used in the literature of AP for the social media domain. Our results show strong evidence of the usefulness of our proposal, which outperforms most of the reference approaches. In fact, the proposed methodology obtained the highest performance in the PAN author profiling competition for two consecutive years [33,32]. In addition to its high-performance capabilities, the proposed approach offer other benefits, namely efficiency, language-independence and interpretability. These aspects of the proposed representation are also evaluated and compared to relevant works.

[1] One should note that this assumption is in fact the underlying hypothesis of classification models: classifiers aim at finding commonalities among instances that belong to a same class.