Towards Multiple Identity Detection in Social Networks

Kahina Gani
Université Blaise Pascal
Clermont-Ferrand 2, LIMOS
UMR 6158, France
gani@isima.fr

Hakim Hacid
Bell Labs, Nozay - France
hakim.hacid@alcatel-lucent.com

Ryan Skraba
Bell Labs, Nozay - France
ryan.skraba@alcatel-lucent.com

ABSTRACT

In this paper we discuss a piece of work which intends to provide some insights regarding the resolution of the hard problem of multiple identities detection. Based on hypothesis that each person is unique and identifiable whether in its writing style or social behavior, we propose a Framework relying on machine learning models and a deep analysis of social interactions, towards such detection.

Categories and Subject Descriptors

H.3.5 [**Information Systems**]: Information storage and retrieval—*On-line Information Services*

General Terms

Experimentation

Keywords

Social Networks, Authorship analysis, multiple identities

1. INTRODUCTION

Social networks offer the possibility for users to communicate behind one or multiple identities, whether their own, someone else's, or fantasy. Identity theft of a third party, including publishing a "fake profile", can have serious consequences, by taking advantage of the reputation of a celebrity or a brand, or to ridicule someone in the public eye. A more serious use of multiple identities is to engage in cyber-crime or fraud. In this context, investigators generally monitor the behavior of users by going through the suspected identities manually or with rudimentary tools. This is an arduous process since: (i) there is a large number of users generally involved in the network, (ii) several combinations of users are possible, (iii) the life time and the dynamics related to the interactions of the user, (iv) the diversity in the profiles an identity interacts with and, (v) if a user uses different identities for an illegal operation, she/he will try to hide as much details as possible to prevent leaving traces. The difficulty of the investigator is to identify such identities in a context where structural information are not enough reliable to make a decision.

We define the problem we are dealing with as follows: *having a large set of social networks users and their associated records of interactions, how to* ***support*** *analysts (e.g., criminal investigators) in* ***targeting the search*** *for different identities corresponding to a specific user.* The idea is then to transform the problem of finding different identities corresponding to the same user to a task manageable by a human brain by building reduced sets of users that the human brain can easily analyze and verify. It is mandatory that these reduced groups have to be built according to reasonable features which have to clearly capture the phenomena of multiple identities. To provide such support, our working hypothesis is the following: *Even if a social network user decides to create several identities for different purposes, and even if she/he tries to hide the relations between the different identities she/he creates, a link may be established between those identities thanks to latent habits of the user which may manifest from, e.g. his writing styles, interests, temporal behaviors, etc. which construct the* ***signature*** *of the identity of the user.* The idea is then to search for a possible representation of the signature of each user and then find a grouping strategy which may reveal a plausible link between the identities.

Existing work in this domain of *authorship analysis*[1] can be divided into three categories: (i) *authorship identification*, which compares anonymous texts against those belonging to identified entries[2], (ii) *authorship characterization*, which attempts to determine the socio-linguistic profile of the characteristics of the author[2], and (iii) *similarity detection*, which compares texts, for example, in plagiarism detection[3]. As part of our study we considered a similarity detection approach. To the best of our knowledge, there is no work that provides authorship analysis using the special characteristics that may be found in social networks (links, activities, tags, etc.). In fact, no work has tackled the problem, in the context of social networks. This is mainly due to the complexity of the problem, the huge amounts of data and the lack of labeled data.

2. OUR APPROACH

Our proposed framework for the detection of multiple identities in social networks (illustrated in Figure 1) uses the methods of authorship analysis combined with machine learning techniques, specifically a k-means and a Kohonen map[4]. It consists of three layers: *Representation Space* for choosing and extracting features from content and relationships, *Learning layer* for applying clustering methods to group similar identities, and *Validation* which is the final step where human intervention makes decisions on which identities are managed by the same author.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

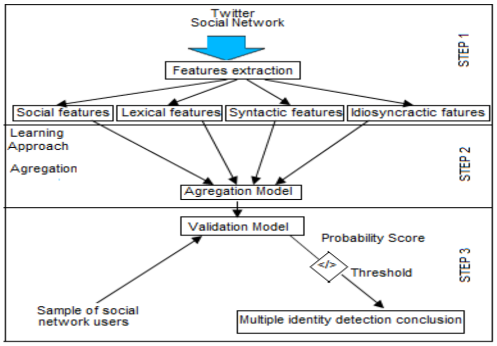

Figure 1: Methodology of the proposed framework

The features chosen in the **Representation Space** significantly affect the performance of our search for multiple identities. Our framework manages the following *lexical features*: average words length, average number words per message, average message length, ratio of uppercase letters, average number of short words, ratio of short words per message, ratio, standard deviation and variance of special characters, and ratio, standard deviation and variance of emoticons. The *syntactic features* include grammatical analysis to determine noun usage (singular vs. plural and common vs. proper). We also monitor the use of 363 function words and 8 punctuation marks. Among the *idiosyncratic features* we use are intentional and unintentional spelling errors, using a dictionary and TF-IDF. This is important since the nature of social networks encourages the creation of new ways to express oneself in short-hand. An important contribution of our work includes *social features*, composed of *Activity* (the ratio of message type sent by an identity over a given time interval, which gives us the degree of participation of the user in the network, and *Topology* which characterizes the social neighborhood of a specific user, such as (in Twitter), links of friendship, followers or followings, as well as the density index of each identity.

As opposed to existing approaches that have a set of known authors being to unattributed work by supervised learning, our framework uses an unsupervised approach and especially two specific clustering algorithms: k-means [5] and Kohonen Maps [6], used to group the data into clusters of high similarity. In the **Learning** layer, we have decided to apply machine-learning techniques on separate subsets of features (that we call subspace from now on) and merge the results, for the following reasons: (i) our initial hypothesis is that a user may intentionally or unintentionally disguise a subset of features between multiple identities, but cannot control all the features such as idiosyncratic features, and (ii) smaller feature sets provide better performance in machine learning tasks. The merge, or aggregation, of all the clusters obtained in the different sub-spaces is performed by calculating a minimum frequency of co-occurrence of each identity in the sub-space clusters. This provides much smaller clusters than in each subspace separately, facilitating the analysis of the grouped identities.

3. PRELIMINARY RESULTS

For applying the previous steps, we retrieved messages of over one million accounts of Twitter social network, and we studied the features of these messages relatively to the writing styles of the different identities. We obtain different results depending on the subspace used (four subspaces are considered) and the method used (k-Means or Kohonen Map) as illustrated in Figure 2. Clusters having a small number of identities may simplify the work of an investigator since she/he has a reduced set of identities to check. We observe that the two methods provide different groupings, but both of them provide almost the same proportion of clusters having few identities. k-means is a more permissive model, providing a dominant cluster containing most of the identities and other very small clusters.

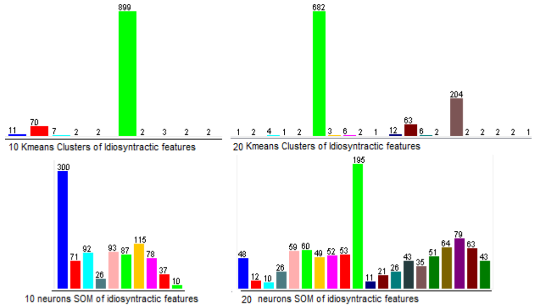

Figure 2: Illustration of the obtained results with 1000 users and a configuration of 10 and 20 clusters.

4. CONCLUSION AND FUTURE WORK

We have presented a framework for grouping similar identities, and supporting the discovery of multiple identities belonging to one author, through three levels: (i) representation space extraction, (ii) learning layer, and (iii) validation. Among the features which are already widely used in research in the field of authorship analysis (lexical, syntactic and idiosyncratic), we have considered features characteristic of social networks (activity and topology). It has been shown that since a user cannot manage an infinite set of identities, the threshold has been set to 20. This needs to be validated by domain experts, but is already an interesting reduction of the search space for authors. As immediate future work, we plan mainly to evaluate the obtained groupings with a larger dataset and directly with experts.

5. REFERENCES

[1] O De Vel. Mining e-mail authorship. volume 30, page 55, 2000.

[2] Malcolm Walter Corney. Analysing e-mail text authorship for forensic purposes. 2003.

[3] F Iqbal, R Hadjidj, B Fung, and M Debbabi. A novel approach of mining write-prints for authorship attribution in e-mail forensics. volume 5, pages S42–S51. Elsevier, 2008.

[4] Jiawei Han. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[5] Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *ACM KDD*, pages 551–556, 2004.

[6] T. Kohonen, M. R. Schroeder, and T. S. Huang, editors. *Self-Organizing Maps.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 3rd edition, 2001.