Reconciliation of Profiles across Social Networks

Master's Thesis Proposal

Nacéra BENNACER, Gianluca QUERCINI

LRI - CentraleSupélec

nacera.bennacer@lri.fr, gianluca.quercini@lri.fr

1 Introduction

Online social networks, such as Facebook, Twitter, LinkedIn, and Weibo, are widely-used platforms that enable key social and professional interactions among millions of individuals across the world.

The data published on social media are a unique trove of information that already proved to be a valuable asset in numerous research domains, especially social sciences [1, 3]. These data are largely noisy, heterogeneous (e.g., graph data and textual content), uncertain (e.g., the opinions of users), sometimes false (e.g., rumors or intentional lies), multilingual and often ambiguous (as in the case of toponyms). People tend to make publicly available some personal information, such as their names, family names, locations and birth dates, just to name a few. However, it is not uncommon that individuals create multiple profiles in several social networks, each containing partially overlapping sets of personal information. Matching those different profiles allows to create a global profile that gives a holistic view of the information of an individual.

The objective of this proposal is to investigate efficient methods to detect profiles referring to the same individuals across multiple social networks in order to integrate all the information regarding the individuals themselves.

2 Contributions

In a previous work [2], we proposed and evaluated an approach that uses a set of rules to match (or, reconcile) profiles based on a limited number of information disclosed in the profiles, such as *nickname*, the *name* and the *geographic location*. One key problem of this approach is that it trusts the information found in the profiles and makes no attempt to verify their correctness and thus detect false/outdated information. Also, this approach is applicable to social networks that have an abundant amount of public information, which is rarely the case.

In the context of a Master's thesis, our goal is to propose, implement and evaluate a new approach that should be :

- robust to false and incomplete information;
- applicable to social networks that have scarcity of publicly available data; and
- scalable.

The objective is to achieve a parallel implementation of the approach by using Big Data technology such as *Hadoop*.

3 Organization and Expected Outcome

The master's thesis will span a period over 5-6 months during which the selected student will be based at the Laboratoire de Recherche en Informatique (LRI). The work, that will be joinly supervised by Nacéra BENNACE, associate professor at CentraleSupélec and Gianluca QUERCINI, assistant professor at CentraleSupélec, will be organized as follows:

- **Month 1.** Thorough bibliographic survey on the topic to select the most important research work in the field and, possibly, to evaluate existing approaches.
- **Month 2 - 3.** Definition of an approach that responds to the criteria cited above.
- **Month 3 - 4.** Implementation of the approach and evaluation on a selected dataset. A comparison against existing approaches will be included.
- **Month 5.** Preparation of the thesis and of a research paper to submit to an international conference and/or a peer-reviews journal.

References

[1] Michelson, M., and Macskassy, S. A. Discovering Users' Topics of Interest on Twitter: a First Look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data* (2010), ACM, pp. 73–80.

[2] Quercini, G., Bennacer, N., Ghufran, M., and Jipmo, C. N. LIAISON: reconciLIAtion of Individuals Profiles Across SOcial Networks. *Advances in Knowledge Discovery and Management* (2016), 229–253.

[3] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al. Personality, Gender, and Age in the Language of Social Media: The Open-vocabulary Approach. *PloS one 8*, 9 (2013), e73791.