Web Science and Engineering 2015

Homework 6

Eric Camellini
4494164
e.camellini@student.tudelft.nl

ABSTRACT

In this homework I analyze how Web Science can be used to gather useful information in heterogeneous fields. In particular I consider three case studies where Web and Data Science techniques are applied to obtain useful information about migration, personality and epidemics (refer to [4], [1] and [2] respectively). For each one of these works I identify the Research Questions that guided it. Eventually I define new metrics and theoretical concepts that I think could help to understand migration, personality or epidemics, following the vision outlined in [3].

1. MIGRATION

The Research Questions of [4] are:

- Can exploiting digital records help to quantify international migration flows?
- Can e-mail messages be used to estimate age and gender-specific global migration rates?
- What is the most accurate way to estimate the geographic location from where an e-mail message was sent?
- How can Internet penetration rates by age, gender and country influence a study that wants to estimate age, gender and country related information using e-mail messages?
- What is the best way to correct the bias that derives from differences in Internet penetration rates?
- How can migration be defined using e-mail messages sending locations over a certain time window?
- Are people moving across borders more or less than at the end of 2009?
- Are there differential migration trends by age and gender?
- What measures can be used to evaluate changes in mobility over time?

2. PERSONALITY

The Research Questions of [1] are:

- Are some features of a generic Facebook profile correlated with its owner's personality, as measured by the standard Five Factor Model personality questionnaire ?
- Is it possible to determine the personality of a person, as measured by the standard Five Factor Model, using specific features extracted from his/her Facebook profile, rather than a personality questionnaire ? If yes, what are the best features that can be used to infer each one of the five factors in the model ? And which is the statistic/machine learning method that results in the highest accuracy?
- Are Opennes and Neuroticism positively correlated with the number of status updates, photos, groups and "likes" of an individual?
- Is conscientiousness negatively correlated with all aspects of Facebook use (i.e. number of friends, likes, photos, etc.) ?
- Is Extraversion positively correlated with all aspects of Facebook use ?
- Is Agreableness positively correlated with the number of Facebook friends, groups and "likes" ?

3. EPIDEMICS

The Research Questions of [2] are:

- Can additional sources of information, such as social media streams, provide complements to the traditional epidemic intelligence mechanism?
- Could it be possible to generate an early warning signal for an epidemics, before well established systems, by only tracking Twitter?
- Could it be possible reduce the dimensionality of Tweets in order to filter information items, by representing them in a low-dimensional space?

- What is the best approach to extract from unstructured textual content, such as Twitter messages, information that can help the investigation of an outbreak?

- What is the best approach to rank small elements of textual content, such as Twitter messages, to support the investigation of an outbreak ?

4. NEW METRICS AND CONCEPTS

In this Section I propose new metrics and new theoretical concepts that can help to investigate and understand the applications from [4], [1], and [2], taking cue from the vision outlined in [3].

Useful metrics for a better understanding of the three listed topics could be built around the mobile access to the Web.

For what concerns migration, it could be useful to identify *travel related posts* on Social Networks: a *travel related post* could be defined as a post that contains information related to travelling, or that is sent from a Mobile device while the user is travelling to another country. Analyzing, for example, the flow of this kind of posts could help predicting the migration flows. Furthermore, more work could be done on the content of these posts: it could be used to distinguish between the different reasons that motivate the migration (e.g. study, work or holiday) or to determine the mood of the users while migrating. Eventually, the mood information could also be used to make some considerations about the social problems related to migration. Extracting travel and mood related information from Social Network posts can be done using the known information retrieval techniques for text mining.

Regarding personality, an extension of the work that has been done could be to use data from the Social Networks together with data related to the mobile access to the Web, in order to improve the prediction of the users' personality. For example, trying to determine which are the posts published by a user during a social activity with other users and checking if this user prefers to post about that activity while doing it or after its end (e.g. at home, when he is alone). A new correlation could be, for example, that a user that uses the Smartphone a lot during an activity with friends is probably less Extraverse than a user that prefers to publish the posts when the activity is over, and other correlations could be found with the other personality factors. Another metric could be taking into account also the personality profile of the other users that participate to the activity: these can be found using the localization functions of the Mobile device, checking the location of the posts on social networks and using information about friends tagged in these posts or about other people that are there at the same time.

Mobile access to the web could also be used to gather information about epidemic intelligence, and especially about the spreading of an epidemic. Useful information could be extracted by identifying subjects that carry the illness and analysing their Mobile posts on Social Networks: as described for the personality (in the previous paragraph), in this way is possible to extract information about their interaction with other users. In this case it means that all these people that interact with the ill subject have a high probability of being infected.

All these Mobile access related metrics can have even more effect in a scenario where also other kind of mobile related data are considered, such as data produced by wearable devices, and can be named as *number of interactions*, *kind of interactions* or, in general, *users interaction metrics* or *users context metrics*. This vision can be seen a a shift from focusing on *what do users post* to focusing on *what is the context of the users while posting, and what do they post?*, and it is possible to answer this question because of the shift in the way in which users access to the Web: Mobile access is becoming the dominant mode.

The described concepts are a clear example on how what users can do with the Web-related technology reflects on what researchers can do with it, in the Web Science field.

References

[1] Y. Bachrach, M. Kosinski, T. Graepel, P. Kohli, and D. Stillwell. Personality and patterns of facebook usage. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 24–32, New York, NY, USA, 2012. ACM.

[2] E. Diaz-Aviles and A. Stewart. Tracking twitter for epidemic intelligence: Case study: Ehec/hus outbreak in germany, 2011. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 82–85, New York, NY, USA, 2012. ACM.

[3] B. Shneiderman. Web science: a provocative invitation to computer science. *Communications of the ACM*, 50(6):25–27, 2007.

[4] E. Zagheni and I. Weber. You are where you e-mail: Using e-mail data to estimate international migration rates. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 348–351, New York, NY, USA, 2012. ACM.