Statistical Relational Learning Towards Modelling Social Media Users

Golnoosh Farnadi
Dept. of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium
Dept. of Computer Science, Katholieke Universiteit Leuven, Belgium
Golnoosh.Farnadi@ugent.be

1 Motivation
Nowadays web users actively generate content on different social media platforms. The large number of users requiring personalized services creates a unique opportunity for researchers to explore user modelling. To distinguish users, recognizing their attributes such as personality, age and gender is essential. To this end, substantial research has been done by utilizing user generated content to recognize user attributes by applying different classification or regression techniques. Among other things, we have inferred the personality traits of Facebook users based on their status updates using three different classifiers [Farnadi et al., 2013]. But as we concluded in [Farnadi et al., 2014b], user attributes are not isolated, as emotions expressed in Facebook status updates are related to age, gender and personality of the authors. Using multivariate regression or multi-target classification techniques is one approach to leverage these dependencies in the user attribute learning process. For example, to employ the dependencies between different personality traits, we applied five multivariate regression techniques to infer the personality of YouTube vloggers [Farnadi et al., 2014c].

The above mentioned techniques are powerful types of machine learning approaches, however they only partially model social media users. Users communicate directly as friends, or indirectly, by liking others content. These types of interaction between users are a valuable source but modelling them with the traditional machine learning approaches is challenging. Furthermore, user generated content in social media comes in different modalities such as visual and textual content, whereas different pieces of evidence might contradict each other. Moreover, in extracting features from each source, a reasonable amount of noise is expected. Hence, using the extracted features as a single feature space without considering features’ noise, dependencies or conflicts, reduces the quality of the learning process. To overcome these limitations, we introduce a new statistical relational learning (SRL) framework [Getoor and Taskar, 2007] suitable for modelling social media users, which we call PSLQ [Farnadi et al., 2014a].

2 SRL with Soft quantifiers
PSLQ is the first SRL framework that supports reasoning with soft quantifiers, such as “most” and “a few”. We start with probabilistic soft logic (PSL), an available SRL framework which defines templates for hinge-loss Markov random fields [Bach et al., 2013] and extend it to a new framework with soft quantifiers. Unlike other SRL frameworks whose atoms are Boolean, atoms in PSL can take continuous values in the interval [0, 1], which facilitates analysis of continuous domains such as user behavior in social media. Indeed, in practice user behavior is not always black-and-white. For example, under interpretation I, I(Friend(Bob, Alice)) = 1 and I(Friend(Bob, Chris)) = 0.2, denote that Alice is a close friend of Bob, while Chris is a distant friend. In models for social media it is common to assume that friends are influenced by each other’s behavior, beliefs, and preferences. PSL, similar to other SRL frameworks, uses the existential (∃) and universal (∀) quantifiers from first-order logic to express this dependency. An often cited example in SRL contexts describing smoking behavior among friends is ∃XY Friend(X, Y) → (Smokes(X) ↔ Smokes(Y)) [Richardson and Domingos, 2006]. This formula states that if two people are friends, then either both of them smoke or neither of them. In this case, the probability that a person smokes scales smoothly with the number of friends that smoke. However, many traits of interest might not behave this way, but instead, having a trait only becomes more probable once most or some of one’s friends have that trait as with smoking. Expressing this dependency requires a soft quantifier, which none of the available SRL frameworks allow.

Syntactically, a quantifier expression in PSLQ is of the form: Q(V, F1[V], F2[V]), where Q is a soft quantifier, and F1[V] and F2[V] are formulas containing a variable V. A formula can be an atom as well as a negation, a conjunction or a disjunction of formulas. Formulas are interpreted in Lukasiewicz logic, i.e., for x and y in [0, 1] (the indicates the relaxation over Boolean values): x ∧ y = max(0, x + y − 1).

Figure 1: Example of “most” and “a few” mappings
$x \lor y = \min(x + y, 1)$ and $\neg x = 1 - x$. Similar to friendship, smoking behavior is represented by varying degrees: for example, Chris might be a heavy smoker, while Alice might be only a light smoker. All these degrees can and should be taken into account when computing the truth degree of statements such as “a few friends of Bob smoke” and “most friends of Bob smoke”. To this end, we define the semantics of a quantifier expression based on the approach of [Zadeh, 1983]. Thus, we define the truth value of $Q(X, Trusts(A, X), Trusts(C, X)) \rightarrow Trusts(A, C)$, which indicates that $A$ trusts $C$ to a degree that a few trustees of $A$ trust $C$. Experimental results show that using soft quantifiers not only expands the expressivity of the model, but also increases the accuracy of the inferred results (Table 1).

### 3 Future directions

The work that we presented here can be extended in the following three directions: (1) Besides social trust, many other AI applications could benefit from the use of soft quantifiers. Modelling users in social media and inferring their attributes with a PSL model is a promising direction for our future work. (2) We defined the semantics of a quantifier expression using the approach of Zadeh. Studying other approaches for quantifiers and their complexity of integrating them into SRL frameworks is a direction for our future work. (3) Designing a suitable PSL model is often time consuming, thus we would like to extend the capability of PSL to learn the structure of data to automatically generate rules when no or little background knowledge is available. This would also include an automatic way of learning the best quantifier mapping for each quantifier expression in a PSL model.

### References


<table>
<thead>
<tr>
<th>Model</th>
<th>PR+</th>
<th>PR-</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSL</td>
<td>0.9770</td>
<td>0.4457</td>
<td>0.8118</td>
</tr>
<tr>
<td>PSL$_Q$</td>
<td>0.9789</td>
<td>0.4670</td>
<td>0.8247</td>
</tr>
</tbody>
</table>

Table 1: Results using 8-fold cross-validation using the Epinions sample with 7,974 trust vs. 701 distrust relations. Values in bold are statistically significant with a rejection threshold of 0.05 using a paired t-test w.r.t. the PSL model.