Review on paper titled "Entity Matching in Online Social Networks"

Social networking is one of the most popular activities on the internet and people tend to make multiple accounts on Online Social Networks(OSN's) to avail various services offered by different Social Networks. Four researchers from US and Israel came together with an aim to extract valuable information from these accounts by linking them together.

Entity Resolution means identifying multiple profiles that belong to the same individual. This cross-referencing might seem simple but has multiple problems involved since users might have different usernames or multiple users might have the same username. For example, there are 17,204 people named John Smith. Thus, normal string matching won't solve it and so we use various machine learning techniques for feature extraction on user's data features.

User Identification and Entity Matching has been addressed by many researchers in the past. Their approaches include: representing each profile as vector and finding their similarity score; comparing all profiles of the first network against all those of the second network to see how much they overlap; and by mapping known auxiliary information on the social network. In this paper, they also used various profile fields for the matching but here they classified a variety of features using supervised machine learning techniques.

The proposed method firstly needs a training set which would be used as an input to the algorithm inducing a classifier. It would then estimate whether the two accounts belong to the same user or not. For the training set, a crawler is used to extract unstructured data from the profiles. After pre-processing, the features are extracted from the corpus which can be categorized as:

1. **Name-based:** To find out similarity between two names. They used algorithms for typographical variations, and character matching algorithms(Ex. Edit distance algorithm) and computed token based similarity; it considers cases where the components of the Profile Name are switched, and phonetic similarity; names which sound similar (Ex. Smith and Smyth).
2. **User-info based**: They extracted 15 features to represent different components of personal information of two users. It includes the distance between their locations - Hometown and current city -, current employer similarity, professional experiences, educational background et cetera. They also computed cosine similarity between two complete profiles. They made sure to give more weightage to the words less frequently repeated in the user profiles since such words add more significance in identifying the profile match.
3. **Network topological based:** This consist of features representing similarity between the network of friends i.e. mutual friends, and mutual friends of friends.

After feature-extraction, the next task is to identify which users are a part of both the networks, which is done by cross-referencing their profiles by their names and labeling them on the basis of multiple feature matching. In the last step, the aim is to build the model for entity matching using machine learning techniques.

This method was actually tested on data from two of the popular Social Networks: Facebook and Xing. The performance calculated was quite high (0.982) probably due to a large number of features considered and the supervised learning algorithms.

Therefore, the Entity-matching is possible by combining multiple features - most importantly, the name-based features. Currently, this research paper was focussed on two networks but the author claims to extend this research to multiple networks. Overall, the research paper was very informative and give clear insights of the various NLP algorithms being used in their model.