Information diffusion on Twitter based on user homophily and semantic relatedness

Sanja Šćepanović
Aalto University, Finland, sanjascepanovic@gmail.com
Phan Thanh Trung *EURECOM, France*
Pan Ben Hui *HKUST System and Media Lab (SyMLab), Hong Kong*
Antti Ylä-Jääski *Department of Computer Science and Engineering, Aalto University School of Science, Finland*

Keywords: *Twitter; information diffusion; semantic relatedness; homophily; social correlation; Wikipedia; knowledge base; Explicit Semantic Analysis; diffusion rate;*

Twitter is a 500 million-user social network and micro-blogging service that is widely used for information dissemination. Tweets have already been applied as an emerging phenomenon in journalism: for early news detection and for covering discussions on a certain topic, as well as in support of political movements. Understanding such a vast system of human knowledge and shared opinions is likely to produce numerous other similar applications. Additionally, research analyses on Twitter dataset promise to provide insights about human behavior and interaction when it comes to information spreading and discussion. The focus of our research, conducted on a sample dataset of public tweets, is: what kind of influence can we infer between social correlation and the diffusion of information? Social correlation between the users is a metric defined by a variety of user profile and behavior features. Before building the metric, we calculate diverse statistic on the dataset to select valuable features for profiling users. Selected features include user profile information, online presence habits and semantic relatedness (SR) of their tweet collections.
Semantic relatedness gives a measure of similarity rate of user tweet contents, and thus the similarity of their preferred tweeting or discussion topics. A novel approach in this paper is to employ Wikipedia data as a knowledge base to calculate semantic relatedness between Twitter users. The existing Explicit Semantic Analysis (ESA) approach is adapted and implemented on an English Wikipedia dataset. An algorithm for obtaining semantic relatedness between two word is developed first, and then that algorithm is employed to calculate semantic relatedness between two documents (tweet collections). The relationship between the social correlation and the diffusion rate of information through replying reveals interesting insights. Under different contexts we observe positive homophily or anti-homophily. In detail, we observe anti-homophily in replying rate when considering user similarity based on semantic relatedness, and positive homophily is exhibited when user similarity features are the online habit and selected profile information. A disadvantage of our work is the sample Twitter dataset, due to its relatively small size and sparse user interaction chains. The future work is to perform our analysis on a larger dataset downloaded to cover specific users or topics and thus to gain insights into other type of user interaction besides replying.