Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 584–593

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

Topical cohesion of communities on Twitter

Guillaume Gadek[a,b,*], Alexandre Pauchet[a], Nicolas Malandain[a], Khaled Khelif[b], Laurent Vercouter[a], Stéphan Brunessaux[b]

[a]*Normandie Univ, INSA Rouen Normandie, LITIS, 76000 Rouen, France*
[b]*Airbus, 78990 Elancourt, France*

Abstract

Nowadays, Online Social Networks (OSN) are commonly used by groups of users to communicate. Members of a family, colleagues, fans of a brand, political groups... There is an increasing demand for a precise identification of these groups, coming from brand monitoring, business intelligence and e-reputation management.

However, a gap can be observed between the communities detected by many data analytics algorithms on OSN, and effective groups existing in real life: the detected communities often lack of meaning and internal semantic cohesion. Most of existing literature on OSN either focuses on the community detection problem in graphs without considering the topic of the messages exchanged, or concentrates exclusively on the messages without taking into account the social links.

In this article, we support the hypothesis that communities extracted on OSN should be topically coherent. We therefore propose a model to represent the groups of interaction on Twitter, the reference on micro-blogging OSN, and two metrics to evaluate the topical cohesion of the detected communities. As an evaluation, we measure the topical cohesion of the groups of users detected by a baseline community detection algorithm.

Keywords: Online social network, community detection, measure of community topical cohesion.

1. Introduction

Online Social Networks (OSN) have taken a huge place in the media and in our lives. The large volume, easy access and rapid propagation of online information make the social networks a perfect example of social interaction between millions of individuals.

To characterise a network, OSN analysis often focuses on high-level coarse-grained characteristics of the network (e.g., degree repartition, statistics, histograms, ...): on Twitter[1], Google+[2], Facebook[3], Sina Weibo[4], Vkontakte[5]. Similarly, the information propagation domain usually focuses either on the information extraction or on the propagation process[6,7], without looking thoroughly at the transmitted information itself. In data analytics, the communities frequently depend on the quantitative values (e.g., communities on Twitter are often detected using the follow relation

[*] Corresponding author.
E-mail address: guillaume.gadek@litislab.eu

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.171

between people, sometimes they do not even consider interaction between users). We are convinced that existing communities on OSN are topically dependent. In other words, detecting accurate communities should take into account the relations between the users but also the topic of the exchanged messages.

Among existing OSN, Twitter has emerged as the reference in micro-blogging platforms with over 500 million public messages per day. On the Twitter network, users can follow each other (new messages are made visible in the timeline), but they can also *retweet* one another, that is, cite a message. One can have a public discussion, *replying* the other's messages. A common way to obtain some attention is to *mention* someone. All these actions can be automatically collected from Twitter. Twitter data collection relies on its APIs: the REST allows to query user information, last tweets, followers lists, ... whereas the Stream API enables the real-time reception of tweets emitted by some selected users or containing one of some selected keywords. Rate-limits are set on the free APIs, which disable the very-high-speed data querying: it is impossible to query and keep up-to-date a follow-links graph, or to receive more than a certain amount of tweets per period of time.

Our groups, or user communities, are defined such that strong social ties exist between their members, and where the members share a common interest in a topic. More in detail, we propose a model of an OSN, Twitter, taking into account Twitter technical constraints (retweets[1], replies, ...) as well as real-life concerns (common interest in a group, strong social links between its members). We propose two metrics, ξ and ρ, to improve the analysis of the detected communities.

The model is applied on real data: a corpus of a few millions of tweets described in Section 4.1. Our first step is the analysis of the texts of the tweets to extract salient topics. Our second step is the detection of communities of users from the graphs of interactions between the users, using a state-of-the-art algorithm. Finally, our third step is the evaluation of the semantic cohesion of the groups to validate our hypothesis.

Section 2 presents the related works in community detection as well as OSN analytics. In Section 3 we expose our model, method and metrics. Section 4 gives the technical details about the experiment, shows the results obtained and highlights some elements to discuss. Finally, Section 5 concludes this article.

2. Related Works

2.1. OSN data analytics

OSN, while taking a growing place in our everyday lives, has recently emerged as an interesting domain of research. The focus is commonly directed towards the influence [1,8,9] (the ability of a user to trigger actions from other users), or towards community detection [10,11].

In the domain of detection of influence and influencers in OSN, *influence* is commonly defined as the ability to *engage*, i.e. to make the other users do something (typically, a RT or a click on a URL). Noordhuis et al. [8] proposed a method to set up a cloud software installation (on Amazon), query Twitter for follow links, build the social graph and get it updated. This allows them to compute a daily PageRank which is their score of influence. However, using the public free API, the queries of follow links are slow, and a very long time (months) is needed to obtain the social graph. Kwak et al. [1] and Lee et al. [12] compared various basic indicators of influence: number of retweets, number of followers, PageRank.

2.2. Community detection algorithms on graphs

Many mathematical algorithms are designed to detect communities in a graph. Cazabet and Amblard [13] proposed a comparison of various state-of-the-art algorithms on large networks (around 400k nodes): CFinder, FastGreedy, InfoMap, Louvain/Blondel [14], iLCD. In this experiment, the Blondel (Louvain) method seems to be, by far, the fastest. Queyroi et al. [15] looked for communities in a named entities co-occurrence graph. They compared different algorithms such as Louvain, Label propagation, Dominant Flows and Markov CLustering for graphs (MCL). Pons and Latapy [16] introduced the Walktrap algorithm, based on random walks in a graph. From these walks they define a distance

[1] denoted RT in the following