Information Processing and Management 49 (2013) 1250–1280

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Nepotistic relationships in Twitter and their impact on rank prestige algorithms

Daniel Gayo-Avello *

Department of Computer Science, University of Oviedo, Edificio de Ciencias, C/Calvo Sotelo s/n, 33007 Oviedo, Spain

ARTICLE INFO

Article history:
Received 6 April 2010
Received in revised form 8 June 2013
Accepted 13 June 2013
Available online 17 July 2013

Keywords:
Social networks
Twitter
Spamming
Graph centrality
Prestige

ABSTRACT

Micro-blogging services such as Twitter allow anyone to publish anything, anytime. Needless to say, many of the available contents can be diminished as babble or spam. However, given the number and diversity of users, some valuable pieces of information should arise from the stream of tweets. Thus, such services can develop into valuable sources of up-to-date information (the so-called real-time web) provided a way to find the most relevant/trustworthy/authoritative users is available. Hence, this makes a highly pertinent question for which graph centrality methods can provide an answer. In this paper the author offers a comprehensive survey of feasible algorithms for ranking users in social networks, he examines their vulnerabilities to linking malpractice in such networks, and suggests an objective criterion against which to compare such algorithms. Additionally, he suggests a first step towards "desensitizing" prestige algorithms against cheating by spammers and other abusive users.

1. Introduction

Twitter is a service which allows users to publish short text messages (tweets) which are shown to other users following the author of the message. In case the author is not protecting his tweets, they appear in the so-called public timeline and they are served as search results in response to user submitted queries. Thus, Twitter can be a source of valuable real-time information and, in fact, several major search engines were including tweets as search results at the moment of this writing.

Given that tweets are published by individual users, ranking them to find the most relevant information is a crucial matter. Indeed, at the moment of this writing, Google seemed to be applying the *PageRank* method to rank Twitter users to that end (Talbot, 2010). Nevertheless, the behavior of different graph centrality methods and their vulnerabilities when confronted with the Twitter user graph, in general, and Twitter spammers in particular, are still little-known.

Thus, this paper aims to shed some light on this particular issue besides providing some recommendations for future research in the area. As it will be later discussed, user ranking in social networks cannot be an end in itself, but a tool to be used for other tasks. Hence, this author is not considering any *a priori* "good" ranking and, instead, he suggests measuring the performance of the different methods on the basis of two desirable features: on one hand presumed relevant users should rank atop – although the actual ordering among them is irrelevant; and, on the other hand, spammers should achieve lower rankings.

The paper is organized as follows. First of all, a comprehensive literature review is provided. It deals with several rank prestige algorithms (some well-known and others lesser-known) which are applicable to social networks; their known vulnerabilities; and some partially related work and proprietary tools outside the scope of this study. In addition to that, Twitter

* Tel.: +34 985 10 43 40.
E-mail address: dani@uniovi.es

0306-4573/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2013.06.003

spam is discussed with a focus on link spam (known as follow spam in Twitter). Then, the different strategies to fight spam in social websites are overviewed. Finally, the research questions are stated and the feasibility of "desensitizing" prestige ranking algorithms against follow spam is analyzed. After that, the experimental framework in which this study was conducted is described: the dataset crawled from Twitter; the elaboration of the subset of relevant and abusive users; and the straightforward nature of the evaluation. Afterwards, results obtained with each of the different ranking methods are discussed along with the implications of the study. Finally, an in-depth analysis of the collected dataset is provided in an appendix: it provides details on the nature of the social network, in addition to some demographical analysis.

2. Literature review

A social network, despite the current association with online services, is any interconnected system whose connections are a product of social relations or interactions among persons or groups. That way, families, companies, groups of friends, or scientific production are social networks.

Social networks can be mathematically modeled as graphs and, thus, graph theory has become inextricably related to social network analysis with a long history of research. Think, for instance, of bibliometric studies that can be traced back to Broadman (1944), Fussler (1987), Gross and Gross (1987), and Lotka (1987), although the work by Garfield (1972) is, with no doubt, the one with the highest impact on the daily life of nowadays scholars. However, it is not our aim to provide a survey on this topic; we recommend the reader interested in social network analysis from a Web mining perspective the corresponding chapters from the excellent books by Chakrabarti (2002) and Liu (2006). Instead, for the purpose of this paper it should be enough to briefly sketch the concepts of *centrality* and *prestige*.

Both centrality and prestige are commonly employed as proxy measures for the more subtle ones of importance, authority, or relevance. Thus, central actors within a social network are those which are very well connected to other actors and/or relatively close to them; this way, there exist several measures of centrality such as degree, closeness, or betweenness centrality.

While centrality measures can be computed for both undirected and directed graphs, prestige requires distinguishing inbound from outbound connections. Thus, prestige is only applicable to directed graphs which, in turn, are the most common when analyzing social networks.

As with centrality, there are several prestige measures such as indegree (the number of inbound connections, e.g. cites, inlinks, or followers), proximity prestige (related to the influence domain of an actor, i.e. the number of nodes directly or indirectly linking to that actor), and rank prestige, where the prestige of a node depends on the respective prestige values of the nodes linking to it – rank prestige is mutually reinforcing and, hence, it requires a series of iterations over the whole network.

Given their importance, and for the sake of clarity, a comparison between the two last prestige measures is provided. Proximity prestige is computed as the mean length of all the shortest paths connecting a given node to the nodes within its influence domain. In other words, proximity prestige measures reach as the mean number of "hops" between a node and all of the nodes linked (directly or indirectly) to it. In contrast, rank prestige takes into account the prestige of nodes linking (directly or indirectly) to a given node – that's why it requires iterative algorithms – and, in some sense, it describes how well connected is a node to other well connected nodes.

Rank prestige is, by far, the most commonly used prestige measure and there exist a number of well-known methods to compute one or another "flavor" of such a measure. In the following subsection we will briefly review the popular *PageRank*, and *HITS* algorithms, in addition to lesser-known (although better targeted at social media) techniques such as *NodeRanking*, *TunkRank*, and *TwitterRank*, besides their weaknesses in different abusive scenarios.

2.1. Rank prestige algorithms

2.1.1. PageRank

PageRank (Page et al., 1998) is, in all probability, one of the best known rank prestige methods because it underlies the Google search engine (Brin & Page, 1998). The *PageRank* algorithm aims to determine a numerical value for each document in the Web, such a value would indicate the "relevance" or "authority" of that given document. That value, also known as *PageRank*, spreads from document to document following the hyperlinks – previously it must be divided by the number of outgoing links. That way, heavily linked documents tend to have larger *PageRank* values, and those documents receiving few links from highly relevant documents (i.e. documents with large *PageRank* values) also tend to have large *PageRank* values.

After iterating a finite (in fact a relatively short) number of steps the algorithm converges; at that moment all the nodes within the graph have got a *PageRank* value by means of which they can be ranked. A notable property of the algorithm is that the global amount of *PageRank* within the graph does not change along the iterations but it just spreads from some nodes to other ones. Thus, if the total amount of *PageRank* in the Web was arbitrarily fixed at 1 we could see the *PageRank* value for a given document as a proxy for the probability of reaching that given document by following links at random (that's why *PageRank* is often described as a *random surfer model*). Such a model is described by Eq. (1), where $PR(p)$ is the *PageRank* value for webpage p, $M(p)$ is the set of webpages linking to p and $L(p)$ is the set of pages linked from p.