Ranking nodes in Social Network Sites using biased PageRank

F. Pedroche[1]

[1] *Institut de Matemàtica Multidisciplinària,*
Universitat Politècnica de València, E-46022 València.
E-mail: pedroche@imm.upv.es.

Abstract

The number of Spanish users of Social Network Sites on the Internet has increased vertiginously in the past year. Now, having a profile in Facebook is a usual fact for most Internet users. Users spend more and more time visiting SNSs. Consequently the number of studies devoted to SNSs has grown enormously. Features that are being investigated are, for example, community structure, detection of leaders, network evolution, etc. In this talk a model to classify the users of an SNS is shown. The method uses the biased PageRank associated to the network. The key concept of the model is the use of the personalization vector. This communication is focused on how to use this personalization vector to model social competencies of the users of the SNSs. The final goal is to rank nodes attending to some social skills of the users. Some numerical examples are shown.

Keywords: Google matrix, PageRank, link analysis, social networking, ranking algorithm, Facebook, Twitter, Myspace

1 Introduction

In a previous conference[1], the author presented some ideas to model a Social Network Site. The model was based on using some kind of biased PageRank associated to the graph representing the network. The examples shown were based mainly on the features of the SNS Myspace.

It is worth recalling that in the beginning of 2008, Myspace was still more popular than Facebook. Since May of 2008 Facebook increased its popularity as indicated by Alexa Traffic Rank[2].

The number of users of SNSs has increased vertiginously. Nowadays Facebook has higher Traffic Rank than Youtube and Myspace, being only overcome by Google.com. Furthermore, if we consider *user time* spend on the internet, the data show that Facebook is the first SNS. For example, in July of 2009 the average internet usage for users in the

[1] ALAMA-2008, held in Vitoria-Gasteiz, September 25-26.

[2] Alexa Traffic Rank measures the popularity of a website. This rank is a three month average of daily visitors and number of unique pages viewed per user per day. See alexa.com

USA was of 4h 39m for Facebook and 2h 31m for Google. In February of 2010 this dates changed to 7h 01m for Facebook and 2h 5m for Google; see Nielsen-online.com.

Regarding users in Spain we recall that Facebook in Spanish was launched in February 2008. In February 2010 there are more than 8.8 millions of users in Spain and more than 400 millions all over the world; see facebooknoticias.com. In a world basis, with data of March, 19, 2010 the average user of Facebook spends nearly 6h per month visiting Facebook, while the user of Myspace spends nearly 1h visiting Myspace. In Spain the average user of Facebook spends 4h 50m using Facebook [4].

From December 2008 to December 2009, Facebook in Spain grew 1147 percent versus the previous year (while Tuenti grew 770 percent) in unique total visitors (comscore release February 25, 2009). Since Febrary 2009 Facebook is the first SNS in Spain attending to the number of total visitors (comscore release April 15, 2009).

This communication addresses some questions that were laid on the table in the ALAMA 2008 conference. In particular, there were the following questions.

- Mathematical analysis of SNSs should be based on real networks?
- Can the searcher Google rank users in an SNS?
- Who other researchers in Spain work in the mathematical analysis of SNSs?

Now we can outline some answers. If we want to model real SNSs we may use real data of networks but we can also use data generated by models that try to imitate the real growth of some complex networks. Some real or generated data sets can be found in, e.g., [8], [9], [10], or [11].

There are some SNSs that are opaque to Google search (e.g., Tuenti) and therefore we, as visitors -not owners- of an SNS, cannot use this searcher to rank people on some SNSs. Regarding the use of the usual PageRank, let us consider the network shown in Figure 1. This network[3] has 14 nodes. It has been generated according to the Barabási-Albert model of *preferential attachment* [1].

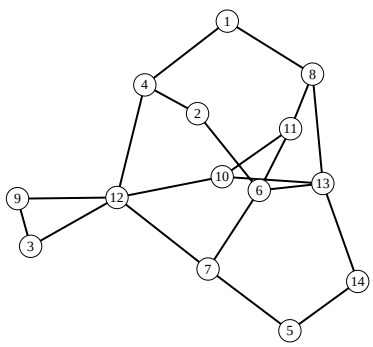

Figure 1: Network from [8]

This network is scale-free and we can assume that represents a small SNS. Note that Nodes 3 and 9 are undistinguishable in terms of the usual PageRank vector.

[3]The corresponding adjacency matrix is available at [8]

Some researchers in Spain are interested in SNSs. For example, A. Arenas (U. Rovira i Virgili), R. Criado (Universidad Rey Juan Carlos), Y. Moreno (Universidad de Zaragoza), E. Moro (U. Carlos III de Madrid), N. Oliver (Telefónica Research and Development), M. Rebollo (U. Politècnica de València), etc.

In the following we recall some definitions from [6] where a method to classify the users of an SNS was presented. This method uses the biased PageRank associated to the network. The key concept of the model is the use of the personalization vector. In this communication we show some examples to illustrate how to use this personalization vector to model social competencies of the users of the SNSs.

2 Definitions

Let $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ be the directed graph representing a Social Network Site. Users are represented by the set of nodes $\mathcal{N} = \{1, 2, \ldots, n\}$ and the hyperlinks are represented by the set of directed links $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$. The link represented by the pair (i, j) belongs to the set $\mathcal{E}$ if and only if there exists a hyperlink connecting node i to node j. In an SNS we assume that each node has at least one outlink; i.e., there are no *dangling nodes*. This is a natural assumption: in an SNS each user has, at least, one friend. Therefore we have $d_i \neq 0$ for all $i \in \mathcal{N}$.

We use the PageRank vector [5] as the main classification tool. Since there are no *dangling nodes* we can define the row stochastic matrix $P = (p_{ij}) \in \mathbb{R}^{n \times n}$, in the form

$$p_{ij} = \begin{cases} d_i^{-1} & \text{if } (i,j) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases} \qquad 1 \leq i, j \leq n.$$

Let $0 < \alpha < 1$ be the so-called damping factor (that we use as $\alpha = 0.85$). Let $\mathbf{e} \in \mathbb{R}^{n \times 1}$ be the vector of all ones and let $\mathbf{v}$ be the personalization (or teleportation) vector, i.e., $\mathbf{v} = (v_i) \in \mathbb{R}^{n \times 1} : v_i > 0$ for all $i \in \mathcal{N}$ and $\mathbf{v}^T \mathbf{e} = 1$. Then the Google matrix is defined as

$$G = \alpha P + (1 - \alpha) \mathbf{e} \mathbf{v}^T,$$

and is an stochastic and primitive (irreducible and aperiodic) matrix [3]. The PageRank vector is defined as the unique left Perron vector of G

$$\pi^T = \pi^T G,$$

with $\pi^T \mathbf{e} = 1$. Denoting $\mathbf{e}_i$ the ith column of the identity matrix of order n, the PageRank of a node i is $\pi_i = \pi^T \mathbf{e}_i$. We call basic PageRank, and denote it by *basic PR* to the vector $\pi(\mathbf{e}/n)$. We recall the following definitions from [6].

Definition 1 Given a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, let $0 < \epsilon < 1$ and let $\mathbf{v}_i = [v_{ij}] \in \mathbb{R}^{n \times 1}$: $v_{ii} = 1 - \epsilon$, $v_{ij} = \epsilon/(n-1)$ if $i \neq j$. For each $i \in \mathcal{N}$, let

$$PR_i = \pi(\mathbf{v}_i).$$

and we denote as $(PR_i)_j$ the jth entry of PR_i.

Definition 2 Given a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ and $0 < \epsilon < 1$, for each node $j \in \mathcal{N}$ we define the *Competitivity interval* $S_C(j)$ as

$$S_C(j) = [\min_{i \in \mathcal{N}} (PR_i)_j, \max_{i \in \mathcal{N}} (PR_i)_j].$$

Definition 3 Given a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, and $0 < \epsilon < 1$ we define the *Competitivity matrix* of the graph, $C = [C_{ji}] \in \mathbb{R}^{n \times 2}$, as follows

$$C_{j,1} = \min_{i \in \mathcal{N}} (PR_i)_j, \qquad C_{j,2} = \max_{i \in \mathcal{N}} (PR_i)_j.$$

Definition 4 Given a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, and $0 < \epsilon < 1$, a *Competitivity group* is a subset of $\mathcal{N}$. Nodes $i \in \mathcal{N}$ and $j \in \mathcal{N}$ belong to the same *Competitivity group* if $S_C(i) \cap S_C(j) \neq \emptyset$.

Definition 5 Given a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, and $0 < \epsilon < 1$, the *Leadership group* is a subset of $\mathcal{N}$. Node $j \in \mathcal{N}$ belongs to the *Leadership group* if, for some $i \in \mathcal{N}$ it holds that $(PR_i)_j \geq (PR_i)_k$ for all $k \neq j$. i.e. for some personalization vector $\mathbf{v}_i$ node j has the greatest PageRank.

3 Modelling Social Competences

Our basic data for analyzing an SNS is the graph associated to it. Yet, It is worth noting that different SNSs have different terms for denoting the connections between users. For instance, in Facebook users have *friends*, while in Twitter users follow other users, becoming *followers*. Note also that, in Facebook, when we consider the relation given by friendship we get an undirected graph, i.e., a connectivity matrix which is symmetric. In Twitter the relation defined by following another user is, in general, non symmetric.

Once the graph is defined our main tool for classifying users is the PageRank vector. We can use the personalization vector to bias the classification to a particular user. To fix ideas let us consider the network in Figure 2, from [2]. In Table 1 we show the basic PageRank obtained for this graph. We also show the corresponding PageRank vectors obtained when using the personalization vectors $\mathbf{v}_3$ and $\mathbf{v}_8$ (see Definition 1). In Table 2 we show the ranking giving by the corresponding PageRank in Table 1.

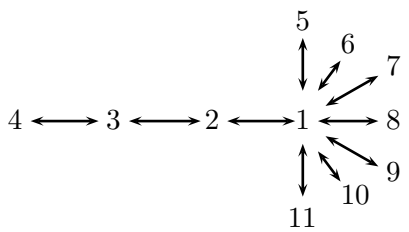

Figure 2: Test network from [2]

basic PR	PR_3	PR_8
0.3700	0.2521	0.4023
0.0972	0.1446	0.0736
0.1042	0.2666	0.0620
0.0579	0.1178	0.0309
0.0530	0.0313	0.0472
0.0530	0.0313	0.0472
0.0530	0.0313	0.0472
0.0530	0.0313	0.1477
0.0530	0.0313	0.0472
0.0530	0.0313	0.0472
0.0530	0.0313	0.0472

Table 1: basic PR, PR_3 and PR_8 for the network of Example 1, with $\epsilon = 0.3$.

ranking with basic PR	ranking with PR_3	ranking with PR_8
1	3	1
3	1	8
2	2	2
4	4	3
5	5	5
6	6	6
7	7	7
8	8	9
9	9	10
10	10	11
11	11	4

Table 2: Node ranking for the network of Example 2 obtained when computing the indicated PageRank, for $\epsilon = 0.3$.

From Table 1 and Table 2 it is clear that the personalization vector can be used to enhance the importance of some nodes. For example, user 3 is the winner when using the appropriate personalization vector. Our idea consists in constructing the personalization vector taking into account some social skills of the user. The idea behind is that the user having high social skills must be prized with an extra PageRank. Our main idea consists in defining a set of parameters to measure the social skills of a user and then we shall construct an appropriate personalization vector that enhances the rank of this user in virtue of its social skills. The weighted vector should be as general as possible to be applicable to different features of the SNSs. The final form of this personalization vector must be tested with the objectives of the managers of the SNS. These ideas are a work in progress.

The managers of an SNS are interested in the number of members they have and in the activity shown by these members. The managers know that the growth of an SNS depends on the activity of the members. When members take part in discussions and interact with other members the result is that other members join the SNS. The quality of the links, as measured by the PageRank, is a good choice for ranking members but we think there are some other features that can incorporate the *activity* of the user. We propose to incorporate this features via the personalization vector.

For example, the following features of Facebook could give an idea of the social skills of a user:

- The number of friends.
- Friends commenting on a comment posted by the user on the own wall.
- Friends commenting on a comment posted by the user on other walls.
- Friends saying they will attend the event invitation posted by the user.
- Friends attending the event invitation posted by the user.
- Friends visiting the *Info* section of the profile of the user.
- etc.

In Figure 3 we show a representation (obtained with the function *spy* of MATLAB) of a matrix corresponding to a Facebook network from a date in September of 2005. This network corresponds to users of Facebook from the California Institute of Technology; See [7] for details[4]. The first analysis that we are currently performing with this data set show that the proposed method is useful to classify the users attending to some features of the users.

4 Conclusions

We have shown a method to classify the users of an SNS. The method is capable of incorporate some features of the users via the personalization vector. The final goal for this model is to incorporate some features related to social skills. We have listed some social skills than could contribute to enhance the importance of a user of an SNS. We have commented preliminary results using test cases.

Acknowledgments

This work is supported by Spanish DGI grant MTM2007-64477.

[4]Data available at $http://people.maths.ox.ac.uk/porterm/data/facebook5.zip$

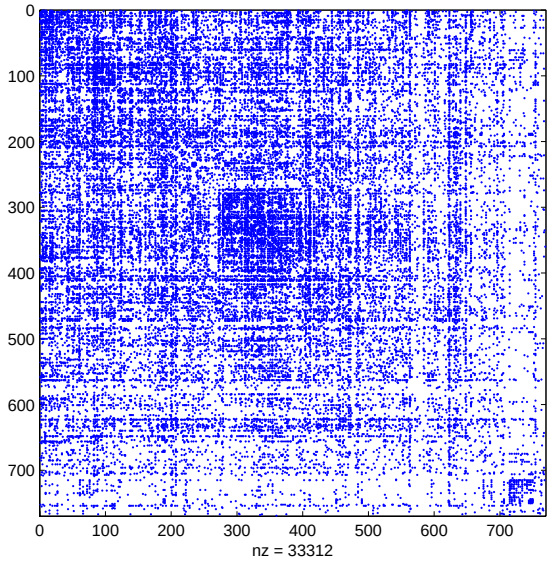

Figure 3: A test matrix from Facebook

Bibliography

[1] R. Albert, A.L. Barabási, Statistical mechanics of complex networks, Rev. Mod. Phys., Vol. 74, No. 1, 2002.

[2] R. Criado, J. Flores, M.I.González-Vasco, J. Pello. Choosing a leader on a complex network. *Journal of Computational and Applied Mathematics, 204 (2007) 10-17.*

[3] A. N. Langville, C. D. Meyer. Google's Pagerank and Beyond: The Science of Search Engine Rankings, Princeton University Press, 2006.

[4] Nielsen Wire. Global Audience Spends Two Hours More a Month on Social Networks than Last Year. March 19, 2010. http://blog.nielsen.com/nielsenwire/global/global-audience-spends-two-hours-more-a-month-on-social-networks-than-last-year/

[5] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project, 1999.

[6] F. Pedroche, Competitivity groups on Social Network Sites, Mathematical and Computer Modelling, 2010, doi:10.1016/j.mcm.2010.02.031. (In press).

[7] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter, Community Structure in Online Collegiate Social Networks, arXiv:0809.0690. (2008)

[8] http://www.infovis-wiki.net/index.php/Social_Network_Generation

[9] http://www-personal.umich.edu/ mejn/netdata/

[10] http://deim.urv.cat/ aarenas/data/welcome.htm

[11] http://www.insna.org/software/public_data.html