P-score: A Publication-based Metric for Academic Productivity

Sabir Ribas
CS Dept, UFMG
Belo Horizonte, Brazil
sabir@dcc.ufmg.br

Berthier Ribeiro-Neto
CS Dept, UFMG & Google Inc
Belo Horizonte, Brazil
berthier@dcc.ufmg.br

Edmundo de Souza e Silva
COPPE, UFRJ
Rio de Janeiro, Brazil
edmundo@land.ufrj.br

Alberto Ueda
CS Dept, UFMG
Belo Horizonte, Brazil
ueda@dcc.ufmg.br

Nivio Ziviani
CS Dept, UFMG & Zunnit Tech
Belo Horizonte, Brazil
nivio@dcc.ufmg.br

ABSTRACT

In this work we propose a metric to assess academic productivity based on publication outputs. We are interested in knowing how well a research group in an area of knowledge is doing relatively to a pre-selected set of reference groups, where each group is composed by academics or researchers. To assess academic productivity we propose a new metric, which we call P-score. Our metric P-score assigns weights to venues using only the publication patterns of selected reference groups. This implies that P-score does not depend on citation-data and thus, that it is simpler to compute particularly in contexts in which citation data is not easily available. Also, preliminary experiments suggest that P-score preserves strong correlation with citation-based metrics.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Academic productivity; Reputation; Publications

1. INTRODUCTION

The assessment of academic productivity usually involves the association of metrics with the researchers or groups of researchers one wants to evaluate. Funding agencies, university officials, and department chairs are examples of entities interested in these metrics, as these have application in a variety of practical situations. There are also cases in which one needs to compare researchers working on a same sub-area of knowledge, some examples are finding review peers, constructing program committees or compiling teams for grants.

Today, the most reliable and complete way to compare researchers is by compiling information on their academic output such as number of publications, citation based metrics, number of undergraduate and graduate students under supervision, number of advised masters and PhD theses, and participation in conferences and in technical committees. Some councils also use extensive surveys to compile qualitative information on features associated with the programs.

However, as compiling this information is not a simple task and takes a long time, it is a common procedure to use just citation data to gain quick insights into the productivity of research groups and academics. But, given that compiling citation counts requires access to the contents of a large pool of publications, which is not always available, new and complementary metrics, such as P-score, are a necessity.

The notion of academic productivity is intrinsically associated with the notion of reputation. And although the concept of reputation lacks on definition, we can see it as a simple property of an individual or group which measures their academic impact in the world and which we can associate metrics with. To measure the reputation of researchers, it is a common procedure to use the publication venues they publish in. Higher the impact of a venue, higher is considered the reputation of the researchers who publish in it. We use this idea of transferring reputation through publications to introduce a new metric called P-score.

2. THE P-SCORE APPROACH

The question we address in this work is: *How to model research groups, researchers and venues to capture the notion of relevance or importance of each, using only information about (i) the relationship of groups and members and (ii) the list of publication records of each member, without using paper contents or citation counts?* Working with this question, we emerged with a metric, which we call P-Score.

2.1 Overview and Assumptions

The basic idea of P-score is to associate a reputation with publication venues based on the publication patterns of a set of *reference groups* of researchers in a given area or sub-area of knowledge. For now we consider that it is possible to select such references, even if it might be controversial.

We assume that the reputation of a research group is strongly influenced by the reputation of its members, which is largely dependent on their publication records. P-score is based on the following assumptions:

1. A researcher or a group member conveys reputation to a venue proportionally to its own reputation.

2. The reputation of a researcher is proportional to the reputation of the venues in which he/she publishes.

Once a reference group in a given area is selected, the reputation of members in this group is transferred to the venues. A Markov chain model can then be built from these ideas.

2.2 Notation and Publication Counts

Before developing the model, we introduce some notation. Table 1 summarizes the notation and definitions used in this work. We use ω and j as indexes for research groups and the venues where they publish, respectively. The research groups used as reputation sources are referred to jointly as the *reference groups*. Consider a chosen set $\mathcal{T}$ of reference groups, and let T be its cardinality. Let $\mathcal{V}$ be the set of all venues v_j where the groups in $\mathcal{T}$ publish, and V the total number of venues in the set $\mathcal{V}$. Members of research group ω publish in subset $\mathcal{V}_\omega \subseteq \mathcal{V}$ with cardinality $V_\omega = |\mathcal{V}_\omega|$.

Table 1: Notation

$\mathcal{T}$	set of reference groups
T	cardinality of $\mathcal{T}$
ω	a research group in $\mathcal{T}$
$\mathcal{V}$	set of venues where the researchers in $\mathcal{T}$ publish
V	cardinality of $\mathcal{V}$
$\mathcal{V}_\omega$	set of venues where the researchers of group ω publish
V_ω	cardinality of $\mathcal{V}_\omega$
v_j	the j^{th} venue where members of a group in $\mathcal{T}$ publishes at
$N(\omega, \mathrm{v}_j)$	total number of distinct papers published by group ω in venue v_j
$N(\mathrm{v}_j)$	total number of papers published in venue v_j
$N(w)$	total number of publications of group ω
$D(v_j)$	number of distinct authors publishing in venue v_j
γ_ω	reputation of group $\omega \in \mathcal{T}$
ν_j	reputation of venue $\mathrm{v}_j \in \mathcal{V}$

We define a function N that counts the papers published by research groups and the papers published at venues. Let $N(\omega, \mathrm{v}_j)$ be the total number of *distinct papers* published by research group ω in venue v_j and let $N(\mathrm{v}_j)$ and $N(w)$ be the total number of papers published in venue v_j and the total number of publications of group ω during the observation period, respectively. That is:

$$N(w) = \sum_{j=1}^{V} N(\omega, \mathrm{v}_j)$$

$$N(\mathrm{v}_j) = \sum_{w=1}^{T} N(\omega, \mathrm{v}_j)$$

2.3 A Markov Model of Reputation

From Assumption 1, the reputation of reference group w is defined as:

$$\gamma_w = \sum_{j=1}^{V} \nu_j \times \alpha_{wj} \quad (1)$$

where

$$\alpha_{wj} = \frac{N(\omega, \mathrm{v}_j)}{N(\mathrm{v}_j)} \quad (2)$$

is the fraction of publications of venue v_j that are from research group ω and V is the number of venues.

Let $D(v_j)$ be the number of distinct authors that publish in venue v_j and T the number of reference groups. From Assumption 2, the reputation of venue v_j is defined as:

$$\nu_j = \sum_{w=1}^{T} \gamma_w \times \beta_{wj} \quad (3)$$

where

$$\beta_{wj} = d \times \frac{N(\omega, \mathrm{v}_j)}{N(w)} + (1-d) \times \frac{D(\mathrm{v}_j)}{\sum_k D(\mathrm{v}_k)} \quad (4)$$

combines the fraction of publications of group ω that are from venue v_j and the fraction of distinct authors that publish in v_j. The intuition for this formulation is venues that receive publications from a small set of authors are most likely to have lower reputation, e.g. local workshops may receive a large amount of publications but the total number of distinct authors tend to be small. The parameter d ($0 \leq d \leq 1$) controls the relative importance between the volume of publications that v_j receives from a group ω and the total number of authors publishing there.

If $d = 1$ then the reputation of the publication venues is totally derived from the reference groups. If $d = 0$ then the reputation of the publication venues is totally derived from the amount of distinct authors (from reference groups or not) publishing there. We noticed that varying d does have an impact on venue weights.

Let $\mathbf{P}$ be a $(T+V) \times (T+V)$ square matrix such that element $p_{mn} = 0$ if either $m, n \leq T$ or $m, n \geq T$. In addition, $p_{mn} = \beta_{m,n-T}$ for $m \leq T, n > T$ and $p_{mn} = \alpha_{m-T,n}$ for $m > T, n \leq T$. Note that, since $\sum_{w=1}^{T} \alpha_{wj} = 1$ for all $1 \leq j \leq V$ and $\sum_{j=1}^{V} \beta_{wj} = 1$ for all $1 \leq w \leq T$ then $\mathbf{P}$ defines a Markov chain. In addition, the Markov chain is periodic and has the following structure:

$$\mathbf{P} = \left[\begin{array}{c|c} \mathbf{0} & \mathbf{P}_{12} \\ \hline \mathbf{P}_{21} & \mathbf{0} \end{array}\right] = \left[\begin{array}{ccc|ccc} 0 & \dots & 0 & \beta_{11} & \dots & \beta_{1V} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \beta_{T1} & \dots & \beta_{TV} \\ \hline \alpha_{11} & \dots & \alpha_{T1} & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{1V} & \dots & \alpha_{TV} & 0 & \dots & 0 \end{array}\right]$$

From decomposition theory, see [2], we can obtain values for ranking the reference groups by solving:

$$\boldsymbol{\gamma} = \boldsymbol{\gamma}\mathbf{P}' \quad (5)$$

where $\mathbf{P}' = \mathbf{P}_{12} \times \mathbf{P}_{21}$ is a stochastic matrix and $\boldsymbol{\gamma} = \langle \gamma_1, \dots, \gamma_T \rangle$. Note that matrix $\mathbf{P}'$ has dimension $T \times T$ only and can be easily solved by standard Markov chain techniques such as the GTH algorithm [1]. Then, from Equation (1) we obtain the reputation of all venues where the reference groups publish.

$$\boldsymbol{\nu} = \boldsymbol{\gamma} \times \mathbf{P}_{12} \quad (6)$$

This vector of venue P-scores can be used to rank authors (or even research groups) one want to compare. But, before continue the development of the P-score model, it is convenient to discuss a small example to illustrate the notation.

2.4 Example

Figure 1 illustrates the Markov chain associated with a small example composed of two reference research groups and three publication venues. In this example, faculty members of Group 1 published a total of six papers, three of

which in venue v_1, two in venue v_2, and one in venue v_3. Venue v_1 got also two papers from faculty of Group 2. Since venue v_1 has a total of five papers from Groups 1 and 2, its reputation is distributed to the two groups proportionally to the number of papers from each. The remaining publication patterns are shown in the figure.

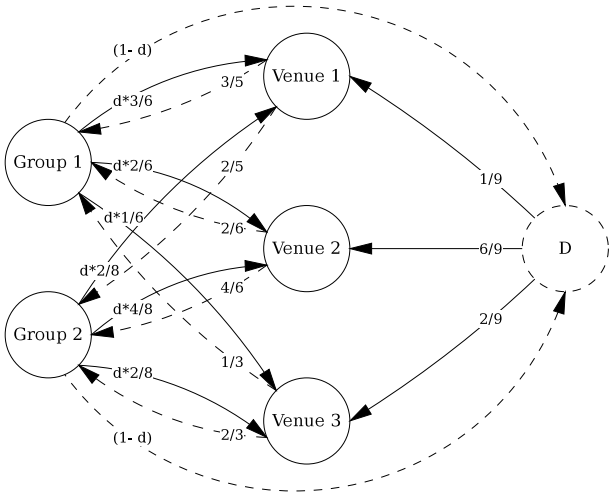

Figure 1: Markov chain for a small example with 2 research groups and 3 venues.

Consider also that we have the number of authors that publish in each venue as an additional information. In our example, assume that venues 1, 2 and 3 receive publications from 10, 60 and 20 distinct authors, respectively. Our intuition is that venues with a larger number of distinct authors are better than venues with a small number of authors (i.e., we penalize venues that are recognized by a few authors). We refer to this effect as the *publication breadth* of the venue. This information is modeled through the *dangling node D* and the parameter $d \in [0, 1]$, which we use to balance the relative importance of publication volume and publication breadth in the model. If $d = 1$ then only publication volume is considered. If $d = 0$ then only publication breadth is considered. For effect of illustration, consider that $d = 1/3$ in our small example of Figure 1. Then, we can write an stochastic transition matrix P as follows:

$$P = \left[\begin{array}{cc|ccc} 0 & 0 & \frac{1}{3}\frac{3}{6} + \frac{2}{3}\frac{1}{9} & \frac{1}{3}\frac{2}{6} + \frac{2}{3}\frac{6}{9} & \frac{1}{3}\frac{1}{6} + \frac{2}{3}\frac{2}{9} \\ 0 & 0 & \frac{1}{3}\frac{2}{8} + \frac{2}{3}\frac{1}{9} & \frac{1}{3}\frac{4}{8} + \frac{2}{3}\frac{6}{9} & \frac{1}{3}\frac{2}{8} + \frac{2}{3}\frac{2}{9} \\ \hline \frac{3}{5} & \frac{2}{5} & 0 & 0 & 0 \\ \frac{2}{6} & \frac{4}{6} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 & 0 \end{array}\right]$$

Given P, we can compute the steady state probabilities associated with each venue to obtain the vector ν of all venues:

$$\nu = \langle 0.189, 0.590, 0.221 \rangle \quad (7)$$

$$= \langle 0.320, 1.000, 0.375 \rangle \quad (8)$$

The values in vector ν are the venue P-scores. In our example, venue v_2 has the highest P-score, followed by v_3, and then by v_1. We remark that the individual values give the relative importance of each venue with respect to v_2.

2.5 Comparing Authors

Once the vector ν of venue P-scores has been computed, we can easily compute a rank $\mathcal{R}$ for each author a in a set of authors A we want to compare as:

$$\mathcal{R}(a \in A) = \frac{S_a}{\max_{i \in A}\{S_i\}} \quad (9)$$

where S_a $(a \in A)$ is a weighted sum of P-scores associated with author a in set A, computed as:

$$S_a = \sum_{j=1}^{V} \nu_j \times N(a, v_j) \quad (10)$$

where ν_j is the weight (or P-score value) of venue v_j according to ν and $N(a, v_j)$ is the total number of publications from author a in venue v_j.

3. DISCUSSION

We have proposed an metric to assess academic productivity, which we call P-score, given it is based just on the *publication* patterns of research groups. The basic idea of P-score is to associate a reputation with publication venues based on the publication patterns of reference groups, composed by researchers, in a given area of knowledge. Although the choice of reference groups can be made by using available citation data, the P-score metric itself does not depend on citation data. It uses just publication records of researchers and research groups, i.e. the papers and the venues where they published in. Preliminary experiments suggest that results have strong correlation with citation-based metrics and yet, have some complementarity to them, something we are further investigating.

ACKNOWLEDGEMENTS

This work was partially sponsored by the Brazilian National Institute of Science and Technology for the Web (MCT/CNPq 573871/2008-6) and the authors' individual grants and scholarships from CNPq, FAPEMIG and FAPERJ.

4. REFERENCES

[1] W. Grassmann, M. Taksar, and D. Heyman. Regenerative analysis and steady state distributions for Markov chains. *Operations Research*, 33(5):1107–1116, 1985.

[2] C. Meyer. Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems. *SIAM Review*, 31(2):240–272, 1989.