A time-aware citation-based model for evaluating scientific products

Francesco Romani
joint work with Gianna M. Del Corso

Dipartimento di Informatica, Università di Pisa, Italy

SMCTools - 19 October 2009

1. The Problem
2. Common metrics
3. The model
4. Introducing time
5. Experimental Results

The Problem

- Number of scientific journals and papers is increasing at an almost exponential rate
- What to read? What to cite? Which journals subscribe? How to evaluate research?
- This burden affects researchers, funding agencies, university administrators, reviewers

Difficult to give an in-depth evaluation of the research

Use indirect indicators of quality

The Problem

Most of “automatic” methods rely on citation analysis

The evaluation of research using citation analysis has weaknesses...

- Is a citation always a trusting vote?
- Data source and coverage
- How do authors choose the papers to cite?

... but also some pros

- Peer review is not always practicable
- There are plausible assumptions underlying the use of citation analysis as a heuristic
- Simple and objective

Notation

We can represent the citation process as a graph and hence as a binary matrix

$$C_{ij} = 1 \text{ iff } p_i \text{ cites } p_j.$$

Assume that receiving a citation is always good!

Common metrics:

Different metrics for different purposes

- Ranking journals - Libraries, scholars for deciding where to publish, ...
- Ranking papers - What to read, what to cite, ...
- Ranking authors - distribution of grants, hiring people, ...

Ranking of Journals

Citation Statistics: Impact Factor, AMS MR, Citeseer,...

Pros: Easy to calculate, time aware, objective, etc.

Cons: Depend on the area.
In the same journal articles with different citation rates.
The ranking provided doesn't always agree with the widely accepted journal's reputation.

Ranking of Journals

PageRank-like techniques: Eigenfactor, SCImago, RedJasper, ...

Based on the idea that not all the citations are equal

Ranking of Journals

Pros: Quality is more important than quantity, metric of prestige, nice mathematical properties

Ranking of Papers

- Relevance of the journal where the paper is published
  - Not all the papers in a journal have the same quality
- Number of citations received
  - Citation gathering can be a very slow process

Ranking of Authors

- Top author if she publishes in “top” journals
- More accurate measures *h*-index, *m*-index, *g*-index, g_1-index.
- Count the number of distinct authors who are citing me.

Our Proposal [ETNA 08, JCAM 09, BMS 09]

- In the classical approach the ranking of journals is based on citations
- The ranking of papers and authors follows from the rank of the journals where the research is published

We proposed an integrated ranking of authors, journals, papers, areas, and institutions

Mutual reinforcement between papers, journals, authors

Our Proposal

We have seen there are different aims of ranking

- Research evaluation by funding agencies
- Hiring in University or in a Industrial context
- Choosing individuals for a research team
- Many others ...

We try to design a tunable method to capture the different needs

General principles

- A paper is important if published in an important journal but also if cited by important papers and authored by important authors
- An author is important if she has important co-authors and has written important papers published in important journals
- A journal is important if collects citations from important journals, publishes important papers by important authors

	J	A	P
J	citation	publication	publication
A	publication	co-authorship	authorship
P	publication	authorship	citation

The model

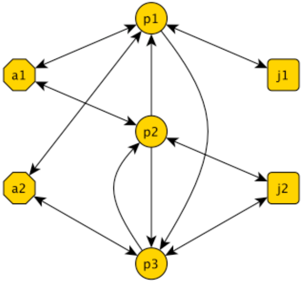

Figure: A graph where we have different nodes for each category. We have three papers, two authors and two journals.

The model

Associate with this graph three matrices, one for each kind of nodes

The journal-paper matrix

$$F(j,p) = 1 \quad \text{if paper } p \text{ is published in journal } j$$

The author-paper

$$K(a,p) = 1 \quad \text{if paper } p \text{ is written by author } a$$

The citation matrix

$$H(r,s) = 1 \quad \text{if paper } r \text{ cites paper } s$$

The model

We can combine these three matrices to obtain the following 3×3 block matrix

$$A = \begin{bmatrix} FHF^T & FK^T & F \\ KF^T & KK^T & K \\ F^T & K^T & H \end{bmatrix}$$

Each block of the matrix expresses the relationship between the subjects belonging to the three classes of *Journals*, *Authors* and *Papers*

First column for the ranking of Journals
Second column for the rank of Authors
Third column for the rank of Papers

The model

In particular

$$A = \begin{bmatrix} FHF^T & FK^T & F \\ KF^T & KK^T & K \\ F^T & K^T & H \end{bmatrix}$$

FHF^T is the cross-citation matrix between journals

KK^T is the co-authorship matrix counting the number of papers between two authors

FK^T stores the number of papers on a given journal of each author

Computing the rank

- We scale the rows of A to obtain a row-stochastic matrix P

$$P\mathbf{e} = \mathbf{e}$$

- The ranking score of subjects is the left eigenvector corresponding to the eigenvalue 1,

$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^T P$$

- The rank value π_j of subject j is the weighted sum of the importances π_i of all the other subjects i which are in relation with j, where the weights are $p_{i,j}$, that is

$$\pi_j = \sum_{i=1}^{N} \pi_i \, p_{ij}.$$

Existence and uniqueness

- A should be irreducible
- Perron-Frobenius theorem guarantees the existence and uniqueness of π
- If A is aperiodic, good convergence properties!

Working with stochastic matrices has advantages from a numerical point of view
The iterative process doesn't need normalization
P describes a Markov chain. If P is irreducible, the chain is ergodic and the stationary distribution π is unique

Introducing time into the model

The model has some problems

- The importance a paper confers to a cited paper does not depend on the time of publication
- In other models the time is enforced considering citations to papers in a restricted time window
- In many disciplines citations occur up to ten years after publication
- Most of the citations are missed

Introducing time into the model

Our idea

- the value of the citations to papers change over the time
- papers that do not receive citations lose importance as time elapses
- old papers that are continuously cited over the years do not lose importance
- recent papers highly cited have a chance to rank high even with lower citation count

Introducing time into the model

Let t_i be the time paper p_i is published

Replace citation matrix H with the matrix $H_T = T H$,

$T = \operatorname{diag}(f(t_i))$, $\quad f : [t_0, t_{\max}] \rightarrow [0, 1]$ is a non increasing function

A good choice is

$$f(t) = \exp(-\alpha\,(t_c - t)),$$

t_c is the current time, α is a constant obtained from the half-life decay time ρ such that $\exp(-\alpha\,\rho) = 1/2$

The decay function

The exponential function is well suited for describing the decay of importance of citations

the value at time t depends only on the time elapsed since the publication!

Easy update of H_T
$t_{c_1} = t_c + \delta$

$$H_{T_1} = \exp(-\alpha\,\delta)\, H_T$$

The question of irreducibility

We have to force irreducibility to guarantee ergodicity!

We introduce a dummy node for each class of subjects

- dummy paper which is cited by all the papers and cites back all the papers except itself
- dummy author who is the author of the dummy paper
- dummy journal is the journal where the dummy paper is published

Probabilistic interpretation The dummy journal is the library, the dummy author is the librarian, the dummy paper is the catalog

How to get P

How to scale the rows of A to get P?
Simplest strategy: divide each row of A by the sum of the entries in the row
Flexible strategy: perform a separate normalization of each block of A

- each block of A is normalized to yield nine row-stochastic matrices
- these matrices are compounded with weights

How to get P

Let $\Gamma = (\gamma_{i,j})$ be a 3×3 row-stochastic matrix, then the matrix

$$P = \left[\begin{array}{ccc} \gamma_{1,1}\, J_J & \gamma_{1,2}\, J_A & \gamma_{1,3}\, J_P \\ \gamma_{2,1}\, A_J & \gamma_{2,2}\, A_A & \gamma_{2,3}\, A_P \\ \gamma_{3,1}\, P_J & \gamma_{3,2}\, P_A & \gamma_{3,3}\, P_P \end{array} \right].$$

The parameters $\gamma_{i,j}$ can be used to tune the role that each class has with respect to the other classes

E.g. $\gamma_{2,3}$ allows to tune how much importance authors give to papers

The weight matrix

a good choice for the weighting matrix is

$$\Gamma = \frac{1}{N} \begin{bmatrix} n_J & n_A & n_P \\ n_J & n_A & n_P \\ n_J & n_A & n_P \end{bmatrix},$$

where $N = n_J + n_A + n_P$ is the size of matrix P

With this weighting strategy, the average value of a paper a journal or an author is the same

Experimental results

The experiments were performed on synthetic data
Real datasets are

- not available and/or usable,
- so incomplete that the characteristics of the bibliographic items do not correspond to real cases

In [DR09] we propose a model for generating synthetic data
The data produced agree with the properties observed on real datasets

Experimental results

- Purpose of the experimentation is to see how time affects the rank
- We produced a dataset with one million of papers, half a million of authors and 5,000 journals, which respects the proportion of the cardinality of the classes in real databases

Experimental results

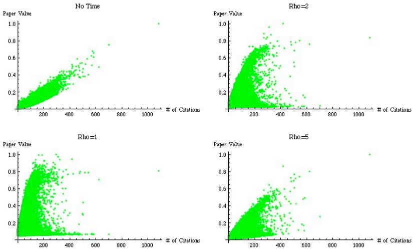

Figure: Dependence of the rank of a paper from the number of received citations.

Experimental results

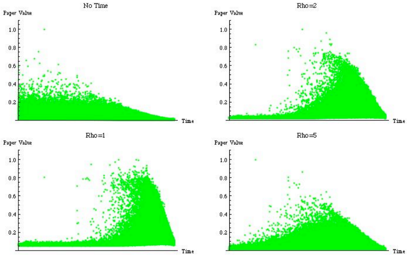

Figure: Rank of papers versus time of publications. We assume a total ordering of papers respect to the time of publication. The top-left plot, depicts the rank of papers with respect to their time of publication.

Experimental results

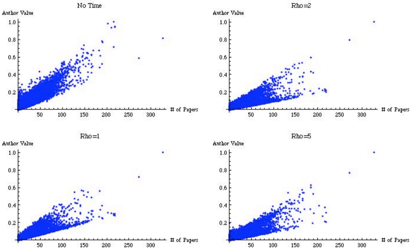

Figure: Dependence of the rank of authors on the number of papers written. We see that with the introduction of the decay in time of the importance of citations we have a more spread plot, even if more productive authors obtain greater rank values.

Experimental results

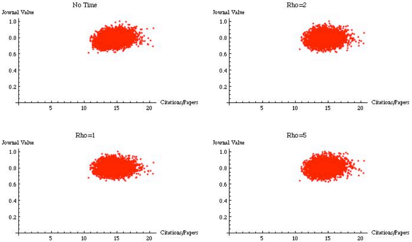

Figure: Dependence of the rank of a journal from the average number of citations received by each paper published on that journal. The situation does not change much with the introduction of the decay in time.

Conclusions

With the intoduction of the time-decay function

- rank is not the effect of a mere citation count
- recent papers are boosted with respect to old papers without many citations in recent years.
- seminal papers published many years ago but still cited do not lose importance because of the fresh citations.

- Interesting theoretical questions need to be further investigated
- Relation between the rank of two chains one obtained from the other by a rank-one perturbation of matrix H_T