Releasing the Power of Digital Metadata

Examining Large Networks of Co-Related Publications

David Tarrant, Les Carr, Terry Payne

UNIVERSITY OF Southampton
School of Electronics and Computer Science

Background

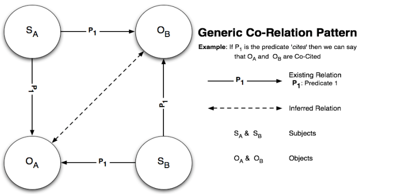
Bibliographic metadata plays a key role in scientific literature, not only to summarise and establish the facts of the publication record, but also to track citations between publications and hence to establish the impact of individual articles within the literature. Currently metrics such as Citation Count and PageRank are widely recognised as suitable methods by which the impact of a publication or article can be calculated. However, as the quantity and accuracy of metadata being made available in digital repositories improves, we can start to realise new metrics for the calculation of impact ranking. In this poster we introduce CoRank, an algorithm which aims to stabalise the impact rank a publication achieves at a point sooner in the publication life cycle than that currently achieved by Citation Count and PageRank.

There are two parts to this work, the main one of which is CoRank, a newer ranking metric which looks at the network of Co-Relations between objects in order to rank those objects. Before we are able to analyse Co-Relations however we have to establish them and this is the job of Co-Pilot. Co-Pilot is an efficient RDF parser which looks for patterns within object records and indexes these ready for use against an ontology. In the scope of this poster we specifically look at the Co-Citation, how this is established and then used in conjunction with CoRank.

Establishing Co-Citations

Large scale repositories of publications currently only index the citations which each publication gives to another. A Co-Citation takes this one level further and looks at the relationship which is established between two articles by a third.

The Co-Citation pattern can be generalised as shown on the right. Here we have two subjects which both share the same relation with two objects. This is an instance of a stronger Co-Relation; the two objects are related with each other twice in this case.


S_A
O_B
O_A
S_B
P_1
P_1
P_1
P_1
Generic Co-Relation Pattern
Example: If P_1 is the predicate 'cites' then we can say that O_A and O_B are Co-Cited
P_1
Existing Relation
P_1: Predicate 1
Inferred Relation
S_A & S_B
Subjects
O_A & O_B
Objects

The Co-Relation has already been used to categorise papers into research areas as well as relating authors in the same manor. We use it here to calculate the impact of research.

CoRank: Speeding up the Publication Life Cycle?

- Current metrics provide a publication life cycle of around 2-3 years, dictated by how long a publications citation graph takes to build.
- The Co-Citation graph of a publication builds 16-20x faster than the citation graph.

PageRank

- Iterative algorithm (linear)
- Operates on **Citation** Graph
- Strength of received score varies based on quality
- Takes approx **2-3 years** to stabalise score

CoRank

- Iterative algorithm (linear)
- Operates on **Co-Citation** Graph
- Strength of received score varies based on quality
- Takes approx **9-14 months** to stabalise score

CoRank

Scaling Factor: For a large number of objects (*V*) this is a very small number. Has the positive effect that a CoRank for a paper (*p*) can never be 0 from the empty graph in the second part of the algorithm.

By traversing the Co-Relation graph the CoRank gained from an object the target object is Co-Related with is the CoRank of that object divided by how many other objects that object is in turn Co-Related with.

$$CR(p) = \frac{1-\alpha}{|V|} + \alpha \sum_{cp_j \in M(cp_i)} \frac{CR(cp_j)}{CL(cp_j)}$$

α	Scaling Factor: Normally set to 0.85.
$CR(x)$	The CoRank of object *x*.
$CL(x)$	The number of items *x* is co-related with.
cp_z	The current co-related object.
$\lvert V \rvert$	The number of objects in the dataset.

The Results Dataset

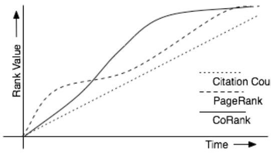
- Citabase Dataset of Physics and Maths Publications (www.citebase.org)
- 32 Snapshots between 2004 & 2007
- 174,786 – 230,076 publications total

citebase

- 500 papers used from first snapshot month for rank analysis and Spearman Correlation.
- 4500 papers used to calculate Spearman Correlation & Paper Age

1 Average Rank Analysis

We look at the average rank by each algorithm of the papers in our dataset. Here we can see that CoRank achieves a stable rank in a much shorter time than both Citation Count and PageRank.



The Spearman Correlation Algorithm

The Spearman Correlation algorithm is used to compare the order of items in 2 dynamic datasets. Here ***d*** represents the difference in the positions of the item in the 2 datasets and ***n*** is the number of items in the dataset.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

2 Dataset Order Comparison

Any new algorithm must be able to produce a similar rank order to current algorithms. The target of CoRank is to do this in a shorter time than that taken by Citation Count and PageRank. Here we made our target algorithm Citation Count and compared the CoRank and PageRank algorithms to it. By using the Spearman Correlation algorithm (above) we show that CoRank does not perform very well, however an improved version called CoRank$_{time}$ (left) does.

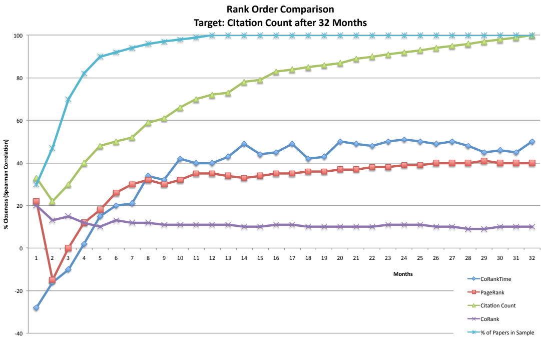
From CoRank to CoRank$_{time}$

CoRank / The Age of the Co-Related Object

CoRank does not perform very well when looking at a dataset order comparison. This is due to the highly cited papers maintaining a high rank by being regularly co-cited with other strong (but old) papers. By dividing the CoRank score by the age of the co-related paper we eliminate this factor.

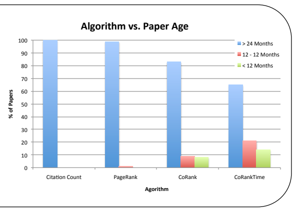
CoRank$_{time}$

$$CR(p) = \frac{1-\alpha}{|V|} + \alpha \sum_{cp_j \in M(cp_i)} \left(\frac{CR(cp_j)}{CL(cp_j)} / (age)\, cp_j + 1 \right)$$



3 Paper Age Analysis

If the publication life cycle is to be speeded up we are looking for the CoRank algorithms to reveal papers which are less than 2 years old. We can see here that both CoRank and the improved CoRank$_{time}$ algorithms achieve this.



More Information:
http://users.ecs.soton.ac.uk/dct05r/publications/
School of Electronics and Computer Science, University of Southampton, Southampton, Hampshire, SO17 1BJ, United Kingdom

David Tarrant: dct05r@ecs.soton.ac.uk
Les Carr: lac@ecs.soton.ac.uk
Terry Payne: trp@ecs.oton.ac.uk