Releasing the Power of Digital Metadata: Examining Large Networks of Co-Related Publications

David Tarrant
School of Electronics and Computer Science
University of Southampton
Southampton, UK
dct05r@ecs.soton.ac.uk

Dr Les Carr
School of Electronics and Computer Science
University of Southampton
Southampton, UK
lac@ecs.soton.ac.uk

Dr Terry Payne
School of Electronics and Computer Science
University of Southampton
Southampton, UK
trp@ecs.soton.ac.uk

ABSTRACT

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. Commercial secondary publishers have typically taken on the role of rekeying, mining and analysing this huge corpus of linked data, but as the primary literature has moved to the world of the digital repository, this task is now undertaken by new services such as Citeseer, Citebase or Google Scholar. As institutional and subject-based repositories proliferate and Open Access mandates increase, more of the literature will become openly available in well managed data islands containing a much greater amount of detailed bibliometric metadata in formats such as RDF. Through the use of efficient extraction and inference techniques, complex relations between data items can be established. In this paper we explain the importance of the co-relation in enabling new techniques to rate the impact of a paper or author within a large corpus of publications.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]; E.2 [**Data Storage Representations**]; G.3 [**Probability and Statistics**]

General Terms

Algorithms, Performance

Keywords

Archiving, Evaluation Methodologies, Metadata, Qualitative Studies, Pattern Templates

1. INTRODUCTION

Bibliometric techniques have emerged as an important mechanism to identify the significance of articles from the literature, and by extension, the quality of the work described. Metrics including Citation Count and PageRank enable the impact ranking of an article to be calculated. In the current publication life cycle model, an articles impact factor using these metrics takes around three years to stabilise. In this poster we look at how Web2.0 and the Semantic Web have influenced the quantity and accuracy of metadata now being made available in digital libraries and how this data can be processed in order to speed up the publication life cycle. By taking the generic "Co-Relation" pattern we look specifically at how the "Co-Citation" graph becomes much larger than that of the citation graph in a much shorter time. We then apply "CoRank", our new ranking algorithm to this graph and find that we can achieve a 50% improvement in time taken from a publications impact factor to stabilised.

JCDL'08, June 16–20, 2008, Pittsburgh, Pennsylvania, USA.
ACM 978-1-59593-998-2/08/06.

2. CORANK

The CoRank algorithm is a logical step beyond PageRank and utilises larger network graphs constructed from the co-relations. Equation 1 outlines the core CoRank algorithm which is run iteratively over each of the publications within out dataset. Like PageRank, the initial CoRank value is set to $1/|V|$ where $|V|$ is approximately the number of publications which exists in the dataset. The CoRank of a paper (p) is generated by taking the CoRank ($CR(cp_j)$) of each paper p is co-cited with and dividing this by the number of papers ($CL(cp_j)$) this paper (cp_j) is co-cited with.

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

From this initial algorithm a series of logical improvement were made before it was found that the best performing algorithm (which is presented in the poster) involved added a factor based upon the age of the co-citation to CoRank (Equation 2).

$$CR(p) = \frac{1- \propto}{|V|} + \propto \sum_{cp_j \in M(cp_i)} \left(\frac{CR\,(cp_j)}{CL\,(cp_j)} / \,(age)\, cp_j + 1 \right) \qquad (2)$$

3. SUMMARY

Extracting the "Co-Relations" which exist within a publication network enables the application of new bibliometrics to these large network graphs. Through the iterative improvements in CoRank, a version of the algorithm has been found which maintains the high impact of well established publications whilst also revealing a number of more recent publications in half the normal publication lifecycle.
More Information:
http://users.ecs.soton.ac.uk/dct05r/publications/