Journal of Informetrics 8 (2014) 546–561

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Contents and time sensitive document ranking of scientific literature

Han Xu*, Eric Martin, Ashesh Mahidadia

School of Computer Science and Engineering, The University of New South Wales, Sydney, Australia

ARTICLE INFO

Article history:
Received 18 December 2013
Received in revised form 28 March 2014
Accepted 8 April 2014

Keywords:
Scientific document ranking
PageRank
Latent Dirichlet Allocation
Topical longevity
Information freshness
Local damping factor

ABSTRACT

A new link-based document ranking framework is devised with at its heart, a contents and time sensitive random literature explorer designed to more accurately model the behaviour of readers of scientific documents. In particular, our ranking framework dynamically adjusts its random walk parameters according to both contents and age of encountered documents, thus incorporating the diversity of topics and how they evolve over time into the score of a scientific publication. Our random walk framework results in a ranking of scientific documents which is shown to be more effective in facilitating literature exploration than PageRank measured against a proxy gold standard based on papers' potential usefulness in facilitating later research. One of its many strengths lies in its practical value in reliably retrieving and placing promisingly useful papers at the top of its ranking.

1. Introduction and motivation

The explosive growth of the Internet and the overabundance of data fuel the creation and development of information networks, which constantly poses new challenges for information retrieval. As the searched domains expand, even queries targeted at some niche field retrieve a large volume of potentially relevant information that far exceeds human processing capabilities. Ranking addresses the challenge of information overload by identifying material of the highest "quality" among all "relevant" material; it has become an integral part of virtually any information retrieval system.

Scientific citation networks are a specific type of information network: they consist of academic publications connected by citations that embody the cumulative research endeavours in scientific domains. Researchers are better enabled to make scientific breakthroughs by taking advantage of current knowledge—borrowing from insights of past studies and availing themselves of data gathered and systems developed—, making exploring citation networks crucial in conducting research. However, citation networks are large in scale and dynamic in nature with high concentration of information and intricate interactions among academic entities (e.g., authors, papers, concepts), making them particularly challenging to navigate. For those reasons, the effective exploration of a citation network requires that high-quality work be identified through ranking. Specifically, recognising publications that have the potential to facilitate later research, or publications with *high scientific utility*,[1] is of special interest as they form the most fruitful part of a scientific field and serve as solid starting points to further explore new possibilities. Intuitively, the scientific utility of a paper is not a constant measure but relative to its contents and age: subsequent research is more likely to benefit from a relatively recent work in a highly relevant topical area whose

* Corresponding author. Tel.: +61 2 9385 6917.
E-mail addresses: hanx@cse.unsw.edu.au (H. Xu), emartin@cse.unsw.edu.au (E. Martin), ashesh@cse.unsw.edu.au (A. Mahidadia).

[1] From here onwards we use the terms utility and usefulness interchangeably.

http://dx.doi.org/10.1016/j.joi.2014.04.006
1751-1577/© 2014 Elsevier Ltd. All rights reserved.

scientific merits have not yet been fully exploited. In this paper, we present a new link-based ranking framework aimed at helping researchers in locating work that contains useful information for them to make progress in their own studies. More particularly, we design our ranking framework to account for both the contents and age of a paper, producing a ranking of papers that better reflects their potential scientific utility.

The rest of the paper is organised as follows. Section 2 briefly surveys the field of scientific document ranking. In Section 3, we identify the gaps in the literature on scientific document ranking. We fill those gaps in Section 4, where we present our link-based ranking framework, specifically designed to model human literature explorers more accurately by taking both document contents and age into consideration. Section 5 discusses experimental results. In Section 6, we conclude and point to future directions.

2. Scientific document ranking

Scientific document ranking is a challenging task whose core problem is to quantify the importance of academic publications. Citation count based metrics have a long lineage, tracing back to the pioneering work done by Garfield on citation analysis in the 1970s (Garfield, 1972, 1979), and they are still widely used today. However, citation count has been challenged for being a quantitative measure of the popularity of a scientific document that fails to properly capture qualitative aspects, such as potential scientific impact (Ma, Guan, & Zhao, 2008; Maslov & Redner, 2008; Sayyadi & Getoor, 2009; Walker, Xie, Yan, & Maslov, 2007; Weingart, 2005). The unique challenge of measuring the scientific value of academic publications has fuelled ever-increasing research efforts.

Link-based ranking approaches such as PageRank (Page, Brin, Motwani, & Winograd, 1999) have been remarkably successful in ranking webpages. By recognising hyperlinks (a form of citation) from one page to another as an implicit conveyance of authority, PageRank calculates the prominence of a page using a less democratic vote, taking the quality of the citing pages into account. Inspired by the success of PageRank in Web search and similarities in problem formulation, a plethora of research has been carried out to rank scientific documents thanks to PageRank or variants of it. Chen, Xie, Maslov, and Redner (2007) directly used PageRank to assess the relative importance of publications in Physical Review journals. By identifying outliers with a moderate number of citations but a high PageRank score, exceptional papers are promoted that would otherwise be disregarded in a citation count based ranking. Ma et al. (2008) also directly applied PageRank to citation analysis and compared their results with the traditionally used citation count metrics. Both studies concluded that, at least empirically, PageRank yields qualitative rankings that resonate better with human judgements. More ambitious studies aim at adapting PageRank to avail themselves of the complex interactions among various academic entities (e.g., publications and authors) in citation networks to incorporate more diverse sources of information. Inspired by the mutual reinforcement between authors and papers—high quality papers are written by renowned researchers and prestigious researchers write high-impact articles—, Zhou, Orshanskiy, Zha, and Giles (2007) proposed the Co-Ranking framework that couples two random walks on the heterogeneous network of authors and papers to produce rankings for both entity types. With similar rationale, King, Jha, and Radev (2013) proposed a simple alternative where basic PageRank was used to generate entity type sensitive rankings in a heterogeneous network of authors, papers, venues, institutions, and terms. Another line of research focusses on generating relative rankings of papers in the "domain frontier". PageRank in its original formulation fails this task as it is negatively biased against young papers that have not yet been given enough exposure to attract citations. To address this issue, some studies proposed to adapt PageRank and take time into account to promote the ranking of recent papers. Walker et al. (2007) designed CiteRank as a random surfer visiting papers to which the assigned probabilities are exponentially discounted as a function of their age. In a similar work, Li, Liu, and Yu (2008) employed both an exponential time decay factor and a trend factor calculated using recent citation time series of a paper to elevate the rankings of new papers. Sayyadi and Getoor (2009) incorporated both a temporal penalty and heterogeneous entities ranking (in this case, authors and papers) into their FutureRank system. They found that time decay plays a much more important role in producing good rankings than the mutual reinforcement between authors and papers.

3. Problem statement

We aim at ranking documents in a citation network to help researchers identify papers of high scientific utility in their field, a paper's usefulness being acknowledged in the kind of incoming citations it receives from later work. A scientific citation network has the same abstract structure as any other directed network, but it is distinctively static in nature: the contents of a document and the references it includes are frozen at the time of publication, imposing a strict temporal constraint on the link structure of a citation network. At any stage, papers only refer to the literature at the time of writing; no pointer to subsequent work can ever be additionally provided. This feature accounts for the fact that a citation network has a strong *ageing characteristic*: first, the temporal constraint causes directed links to point towards progressively older nodes; second, the static nature of a citation network constrains it to evolve with less plasticity than the Web which is constantly updated both in contents and structure. This strong ageing characteristic makes ranking metrics based on citation count unsuitable for the task, although they are traditionally used, as new publications have not been given enough exposure to accumulate citations. Furthermore, it limits the appropriateness of more sophisticated approaches, such as PageRank, which strongly biases its rankings towards older papers. This undesirable bias has the profound implication that rankings produced by PageRank pertain more to "historical importance" and are not reflective of the scientific landscape of a literature at query