Faster Ranking Using Extrapolation Techniques

Muhammad Ali Norozi, Norwegian University of Science and Technology, Norway

ABSTRACT

Extrapolations are techniques in linear algebra that require little additional infrastructure that must be incorporated in the existing query-dependent Link Analysis Ranking (LAR) algorithms. Extrapolations in LAR settings relies on the prior knowledge of the (iterative) process that created the existing data points (iterates) to compute the new (improved) data point, which periodically leads to the desired solution faster than the original method. In this study, the author presents novel approaches using extrapolation techniques to speed-up the convergence of query-dependent iterative methods, link analysis based ranking methods, where hyper-link structures are used to determine relative importance of a document in the network of inter-connections. The author uses the framework defined in HITS and SALSA and proposes the use of different Extrapolation techniques for faster ranking. The paper improves algorithms like HITS and SALSA using Extrapolation techniques. With the proposed approaches it is possible to accelerate the iterative ranking algorithms in terms of reducing the number of iterations and increasing the rate of convergence.

Keywords: Extrapolation Techniques, Link Analysis Ranking, Markov Chain, Ranking Methods, Singular Value Decomposition

INTRODUCTION

Users on the web are expanding from being active information consumers to becoming active information producers. With such a boom in information retrieval and information explosion, there is an ever-increasing demand for access, coverage, quick responses, and relevant results. The huge collection of information inherently entails a loss of performance and efficiency, because it takes time to process (index, cluster, etc), retrieve (query) and keep the information up-to-date in the huge repositories. Thus there is a growing concern about the usability and the interaction time between the user and Information Retrieval (IR) systems. A trade-off between the quality of the results and query response time is mostly considered as an option. In such challenging settings a user query must yield meaningful, manageable and most importantly “relevant” set of results from IR systems in a reasonable time.

The apparent ease with which the users click from document to documents provides a rich source of information which could be used to understand what and where to find the important documents. The semi-structured and diverse collections of documents are held together by the billions of annotated connections called hyperlinks. Analyzing these myriad interconnections between the documents forms the basis for Link Analysis Ranking (LAR).
These analyses will help us identify the proximity and relevance of documents amongst each other, and enables us to find out the social or informational organization of the documents (the sociology of information). We can utilize the contextual exposition of documents to deduce the importance or popularity of the documents in the network, by using the core graph theory concepts and techniques.

The citation structures of the documents contain a wealth of useful but “implicit information”. Through citation structure hundreds and millions of documents can be pulled together into a network of knowledge. Foremost such a structure represents the users’ behaviours and needs. Users on the Web usually discover most relevant and valuable information through the recommendations and references from a good source of information.

One of the main concerns in the link based ranking methods is the convergence to a “good solution” or an equilibrium state. An equilibrium state is a state where system under certain presumptions can declare the set of good results corresponding to user query. Most of the link analysis based ranking models are iterative in nature, they iteratively move towards the required equilibrium state (the good solution). Convergence is a central phenomenon in iterative algorithms (Lay, 1994). In linear algebra, the iterative methods are employed when direct methods would be prohibitively expensive and in some cases impossible even with best possible computing power to find out the actual solution. Essentially the iterative methods such as “power method” (Lay, 1994) provide an approximation to the true solution starting from a seed value. This work deals with the convergence properties and behaviour of the famous query-dependent LAR algorithms (Kleinberg, 1999; Najork, 2007; Borodin, Roberts, Rosenthal, & Tsaparas, 2005; Tsaparas, 2004a; Bianchini, Gori, & Scarselli, 2005).

The major contribution of this work is the improvements primarily in the convergence behaviour of the query dependent LAR algorithms using “careful periodic” applications of extrapolation step during iterations. We have distinctively applied extrapolation techniques to query-dependent LAR algorithms, which was not done before. The parameters are manipulated extensively in the empirical evaluation and hence extracted a very novel performance gain due to extrapolation. We concluded the article with an extensive experimental evaluation. In the study by Kamvar et al. (2003b) they have found an improvement of order 3 at-most due to extrapolation, in PageRank algorithm. By applying extrapolation carefully in the query-dependent algorithms, improvements of order in range (3 – 19) have been discovered in this work, see Table 2 and Appendix in Norozi (2008).

The document is therefore organized as follows; the next section defines the theoretical background and the preliminaries of the problem at hand. Motivation for this paper is also presented in the same section and later the novel idea of extrapolation to speedup the rate of convergence is explained in a reasonable detail in the section Extrapolation. In the Experimental Evaluations section the idea proposed in earlier sections are empirically assessed. In the last section we conclude the study with important results and possible future work.

THEORETICAL BACKGROUND

Link Analysis Ranking (LAR)

The presence of (hyper-) link information clearly augmented a great deal to the characterization of the informative content present in the documents. LAR approaches are intended to resolve some of the intrinsic weaknesses of the content-based Information Retrieval (IR) models. Through the analyses of network of the documents (due to citation structure) LAR approaches bring in a whole new horizon to information retrieval space. The essence of LAR therefore is that the “overall information” of a hyperlink database of documents is not composed of only static “textual information”, but also another, the “hyper” information.

Link Analysis Ranking is the next step from just content-analyses. It involves analyses and understanding of a very huge and jumbled
Related Content

Video Shot Boundary Detection and Scene Segmentation
Hong Lu, Zhenyan Li, Yap-Peng Tan and Xiangyang Xue (2006). *Advances in Image and Video Segmentation* (pp. 188-208).
[www.igi-global.com/chapter/video-shot-boundary-detection-scene/4842?camid=4v1a](www.igi-global.com/chapter/video-shot-boundary-detection-scene/4842?camid=4v1a)

A Modular Framework for Vision-Based Human Computer Interaction
[www.igi-global.com/chapter/modular-framework-vision-based-human/77594?camid=4v1a](www.igi-global.com/chapter/modular-framework-vision-based-human/77594?camid=4v1a)

On the Analysis of Global and Local Robust Image Watermarking Using Zernike Moments
Fuzzy Image Segmentation for Mass Detection in Digital Mammography: Recent Advances and Techniques
www.igi-global.com/chapter/fuzzy-image-segmentation-mass-detection/77574?camid=4v1a