Improving Citation Mining

Muhammad Tanvir Afzal
Inst. f. Information Systems and Computer Media, Graz University of Technology, Austria
mafzal@iicm.edu

Hermann Maurer
Inst. f. Information Systems and Computer Media,Graz University of Technology, Austria
hmaurer@iicm.edu

Wolf-Tilo Balke
Inst. f. Informationssysteme Technische Universität Braunschweig, Germany
balke@ifis.cs.tu-bs.de

Narayanan Kulathuramaiyer
Department of Computing & Software Engineering, Universiti Malaysia Sarawak, Malaysia
nara@fit.unimas.my

Abstract

In recent years the number of citations a paper is receiving is seen more and more (maybe too much so) as an important indicator for the quality of a paper, the quality of researchers, the quality of journals, etc. Based on the number of citations a scholar has received over his lifetime or over the last few years various measures have been introduced. The number of citations (often without counting self-citations or citations from "minor" sources, in whatever way this may be defined), or some measurement based on the number of citations (like the h- or the g-factor) are being used to evaluate scholars; the citation index of a journal (again with a variety of parameters) is seen as measuring the impact of the journal, and hence the importance one assigns to publications there, etc. The number of measurements based on citation numbers is steadily increasing, and their definition has become a science in itself. However, they all rest on finding all relevant citations. Thus, "citation mining tools" used for the ISI Web of Knowledge, the Citeseer citation index, Google scholar or software such as the "publishorperish.com" software based on Google scholar, etc., are the critical starting points for all measurement efforts. In this paper we show that the current citation mining techniques do not discover all relevant citations. We propose a technique that increases accuracy substantially and show numeric evaluations for one typical journal. It is clear that in the absence of very reliable citation mining tools all current measurements based on citation counting should be considered with a grain of salt.

1. Introduction

Citation management is of great importance by providing important input for new research that may otherwise not be possible without "standing on the shoulders of giants". Citations allow authors to refer to past research in a formal and highly structured way [1], to systematically construct a citation network that then serves as a means of valuation for published works.

The citation count, which refers to the number of citations a particular paper receives, is used in evaluating bibliometrics such as the quality of a paper, the quality of researchers, the quality of journals, etc. It has been used for knowledge diffusion studies [2], network studies [3] and in finding relationships between documents [4]. Impact factor measurements, as derived from citation counts have been applied in making important decisions such as hiring, tenure decisions, promotions and the award of grants [5]. As such the determination of precise citation counts is of utmost importance.

Citation mining refers to the process of discovering citation counts. This task in itself is not trivial as it involves extensive text analysis to determine the exact intended citation of authors to published works. Owing to the large number of publications, this task involves a great amount of human effort if done manually. Alternatively, an approach for autonomous citation discovery can be applied. This approach, however, tends to be prone to omissions and mistakes [6]. Fully autonomous citation mining as such has to rely on community

978-1-4244-4615-5/09/$25.00 ©2009 IEEE