A Classification Scheme for Algorithm Citation Function in Scholarly Works

Suppawong Tuarob†, Prasenjit Mitra‡ and C. Lee Giles‡
† Computer Science and Engineering, ‡ Information Sciences and Technology
The Pennsylvania State University
University Park, PA 16802
suppawong@psu.edu, {pmitra, giles}@ist.psu.edu

ABSTRACT

Algorithms are ubiquitous in the computer science literature. A search engine for algorithms has been tested as part of the CiteseerX suite; however, it only retrieves algorithms whose metadata is textually matched with the search query. Such a limitation occurs because a traditional search engine does not have the ability to understand what algorithms are and how they work. Here, we present an initial effort in understanding the semantics of algorithms. Specifically, we identify how an existing algorithm can be used in scholarly works and propose a classification scheme for algorithm function.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]

Keywords
Algorithm Citation, Algorithm Function, Classification

1. INTRODUCTION

An algorithm search engine has been tested as a part of CiteseerX[1, 7, 6]. However, the search is only a text based matching of user queries to algorithm metadata and the results are ranked based on full text TF-IDF. The limitations of the traditional text-based search emphasizes the need for a more semantic understanding of algorithms in scholarly works, such as knowing how researchers utilize the existing algorithms in their work. For example, Walker et al. [9] extended the PageRank algorithm [4] to create a CiteRank algorithm which utilizes the characteristics of citation networks to rank academic publications. Tuarob and Tucker [8] directly used the LDA algorithm [2] to mine product features from tweets.

Here we propose a classification scheme for algorithm citation function in scholarly works. We make an assumption that authors indicate how a previous algorithm is utilized in the context of where the algorithm is cited, which we denote as algorithm citation context. We propose a classification scheme with 9 classes of an algorithm citation function, divided into 3 groups based on the authors’ attitudes. We then manually classify a set of 300 algorithm citation contexts randomly selected from 2,000 papers from CiteseerX and describe the statistics. Discovering the functionality of a cited algorithm could provide an insight into how and where the algorithm is used and could give insight into interesting applications such as algorithm trend discovery and algorithm impact.

2. PROPOSED CLASSIFICATION SCHEME FOR ALGORITHM CITATION FUNCTION

2.1 Classification Scheme

Figure 1 illustrates our proposed scheme for classification of algorithm citation function in scholarly works, along with descriptions. The scheme consists of 9 classes representing possible ways in which an algorithm can be used. The 9 classes can be divided into 3 groups based on the authors’ attitude towards the cited algorithms: favorable, neutral, and critical. We hypothesize that authors are favorable towards algorithms that they choose to incorporate or build upon in their work; whereas they feel critical about algorithms that cannot solve the problems they address. Such criticism is often reflected in discussion about the disadvantages of the cited algorithms.

2.2 Methodology

... What we are planning to do in this paper, we show that this formula (1) can indeed be theoretically justified. Our justification for this formula will use methods motivated by the neural network approach (see, e.g., [10]). We must choose a family of functions, not a single function A, ...
of an algorithm citation context taken from [3]. We manually label 300 algorithm citation contexts randomly selected from CiteseerX, using the scheme in Figure 1.

Figure 2: Distribution of different algorithm citation functions found in 300 randomly selected algorithm citation contexts.

Figure 2 represents the distribution of algorithm citation functions over different classes. Based on the results, most cited algorithms belong to the ‘Mention’ and ‘Argument’ classes. This supports the intuition that authors usually only mention and/or discuss the disadvantages of the existing algorithms to make the readers aware of the motivation and the completeness of their proposed work. We find that authors are mostly respectively 60.99% of the time neutral, 28.34% critical, and 10.67% favorable towards other algorithms. We interpret this to mean that most of the time few algorithms are actually used for further algorithm development.

2.3 Algorithm Citation Function in Different Sections

In this section, we analyze the distribution of the algorithm citation functions as seen in different sections in scholarly documents. The same set of 300 manually labeled algorithm citation contexts are further classified into 6 groups based on the sections in which they appear, including Abstract (ABS), Introduction (INT), Background and Related Works (BCK), Methodology/Experiment/Results and Discussions (RAD), Conclusion (CON), and Other sections (OTHER). Figure 3 illustrates the percentage distribution of each algorithm citation function over different sections.

Figure 3: Distribution of algorithm citation functions over different sections.

3. ACKNOWLEDGMENTS

We gratefully acknowledge partial support from the National Science Foundation (NSF).

4. REFERENCES