Algorithm Citation Context Labeling Instruction

Introduction

An algorithm search engine has been tested as a part of Citeseer^X [1]. The system extracts pseudo-codes along with their metadata from scholarly documents, indexes them with Apache Solr/Lucene[1], and makes them searchable via full text search. However, the search is only a textual 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 emphasize 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. [8] ***extended*** the PageRank algorithm [4] to create a CiteRank algorithm which utilizes the characteristics of citation networks to rank academic publications. Tuarob and Tucker [7] ***used*** the LDA algorithm [2] to mine product features from tweets.

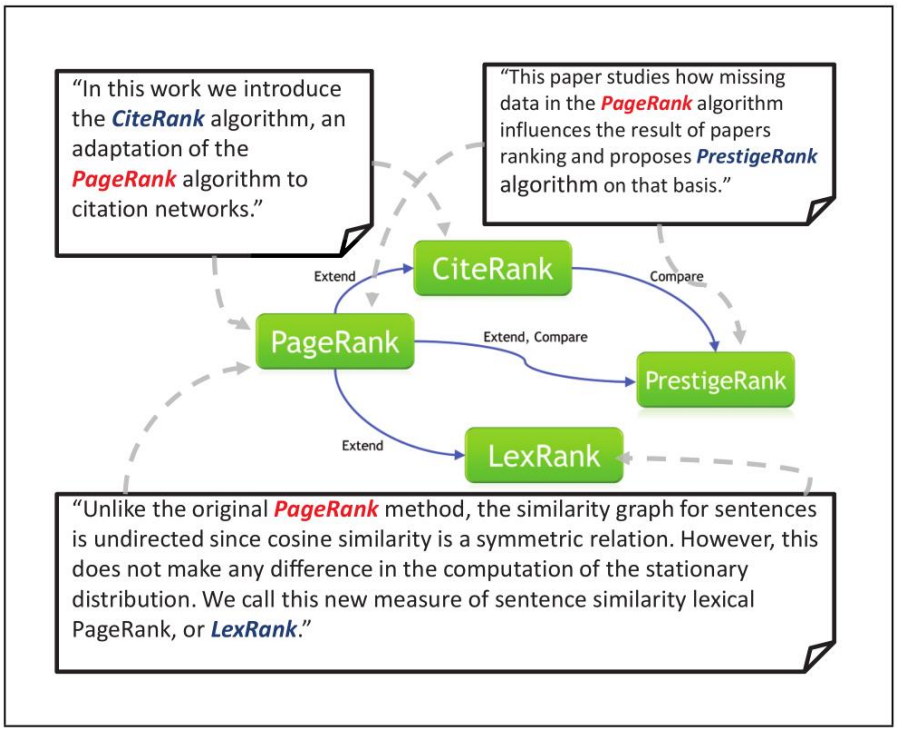

Figure 1: Example of the algorithm citation network

Automatically knowing what an algorithm actually does could shed light on multiple applications such as algorithm recommendation and ranking. Our ongoing work on the algorithm semantic analysis involves studying how algorithms influence each other over time. This study would allow us not only to discover new and influential algorithms, but also to study how existing algorithms are applied in various fields of study. In order to do this, we propose to construct and study the *algorithm citation network*, where each node is an algorithm, and each direct edge represents how an algorithm *uses* existing algorithms. Figure 1 shows a small example of the algorithm citation network, originated from the *PageRank* algorithm [5]. Subsequently, Erkan and Radev proposed *LexRank* which extended *PageRank* [3]. Similarly, Walker et al. introduced *CiteRank* which also extended *PageRank* [8]. Recently, Su et al. proposed *PrestigeRank*, an extension of *PageRank*, and compared the results with both *PageRank* and *CiteRank* [6].

In this work, we propose to study the algorithm evolution using the algorithm citation network generated from a collection of scholarly documents. An algorithm citation network is a heterogeneous direct graph where each node is a paper. For a given citing paper *A* and cited paper *B*, three types of edges are introduced:

USE(A,B) Representing that paper *A uses* one or more algorithms proposed in paper *B*.

EXTEND(A,B) Representing that paper *A* develops a new algorithm that *extends* one or more algorithms proposed in paper *B*.

MENTION(A,B) This simply means that paper *A mentions* one or more algorithms proposed in paper *B*.

We propose to identify the relationship between the citing paper ***A*** and cited paper ***B*** using the algorithm citation contexts found in paper ***A***. An algorithm citation context is a snippet in the citing paper consisting a sentence that the algorithm citation occurs and the sentences that come immediately before and after it. Given the citing paper A and cited paper B, ACCONTEXT (A, B) = {c1, c2, c3, ...} is a set of algorithm citation contexts in the paper ***A***, that cite algorithms in paper ***B***.

Algorithm Citation Context Classification

We make the case that the usage of cited algorithms in the citing paper can be captured in the algorithm citation contexts. For example,

USE(A,B)

> Therefore, it might be desirable to present a set of optimal solutions that the end-user will be able to choose from. We have used Deb?s NSGA- II =-=[13]-=- to generate sets of solutions. NSGA- II is a fast and elitist genetic algorithm frame- work designed for dealing with multi- objective op- timization problems.

EXTEND(A,B)

> We present the Iterative Accelerated A* (IAA*) algorithm for trajectory planning in Sec- tion 3. This algorithm is an extension of the Ac- celerated A* (AA*) algorithm =-=[7]-=-. The orig- inal AA* uses a variable discretization step size to reduce The work was supported by the Federal Aviation Administration (FAA) under project num- ber DTFACT-08-C-00033 and by Czech Ministry o.

MENTION(A,B)

> is refered to [7]. In [6], we pro- posed an algorithm based on the SPLICE technique for speech enhancement. In the same work, a speech detector based on the energy in the bone channel was proposed. In =-=[8]-=-, we proposed an al- gorithm called direct filtering (DF) based on learn- ing mappings in a maximum likelihood framework. However, one drawback with the DF algorithm is the absence of a strong speech mode.

Algorithm Citation Context Labeling Instruction

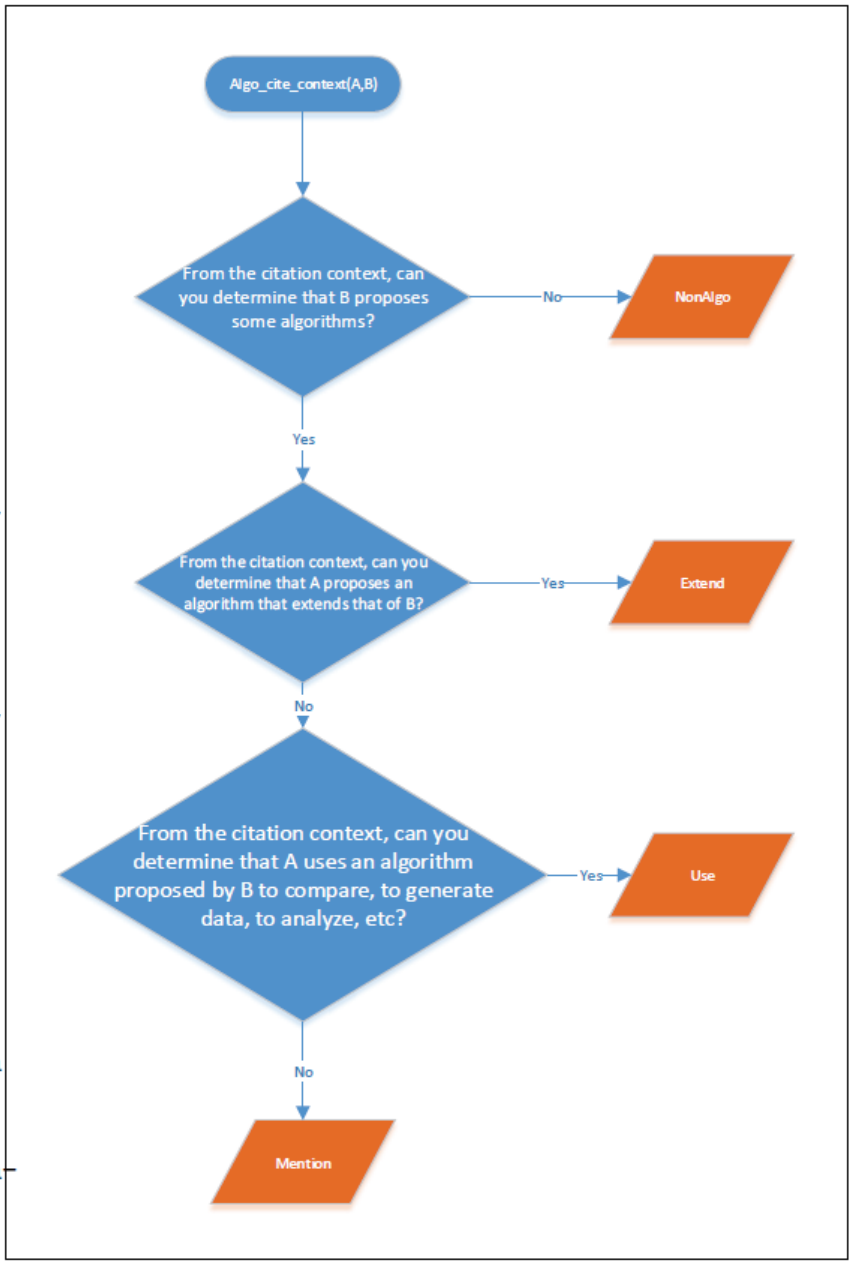

Figure 2: Tagging Instruction

Figure 2 illustrates the step-by-step instructions for tagging a citation context.

1. Identify the focus cited algorithm-proposing paper (B). Its citation is enclosed with =-= and -=-, e.g. =-=[7]-=-, [=-=Click 1995-=-], and =-=[RR98]-=-.
2. If you cannot determine that ***B*** proposes an algorithm (or method, function, scheme, procedure), then tag ***NotAlgo***. Done.
3. Else. If the citation context indicates that the citing paper extends the algorithm proposed in the cited paper, tag ***Extend***. Done.
4. Else. If the citation context indicates that the citing paper uses the algorithm proposed in the cited paper, tag ***Use***. Done.
5. Else. If the citation context indicates that the citing paper merely mentions the algorithm proposed in the cited paper, tag ***Mention***. Done.
6. Else. Tag ***NotSure***, and note your doubts.
7. Note important and interesting characteristics or observations.

Example:

Use	Extend	Mention	NotAlgo	NotSure	context	Note
	1				We present the Iterative Accelerated A* (IAA*) algorithm for trajectory planning in Sec- tion 3. This algorithm is an extension of the Ac- celerated A* (AA*) algorithm =-=[7]-=-. The orig- inal AA* uses a variable discretization step size to reduce The work was supported by the Federal Aviation Administration (FAA) under project num- ber DTFACT-08-C-00033 and by Czech Ministry o.	keyword Extend

Data Analytics

After labeling all the citation context. Please calculate and report the following:

1. Inter-agreement among the 3 raters using Fleiss's Kappa statistics.
2. Frequency of each class.
3. What observations have you observed that can help a computer to automatically classify each citation context?

Deliverables

1. All the labelling files (A, B, C, and combined).
2. A report paper
3. Presentation (either face-to-face or a video).

Tips:

- Honest: Tag your own data as instructed. You must not copy from your friends. You must not guess randomly. The tagged data will be used as training data to train machine learners; hence, the data must reflect facts.
- Focus: Allocate a chunk of 1 hour every day, preferably in the morning. Do not tag the data while feeling drowsy or stressed.
- Consistent: Tag 20-30 samples every day. If you have to skip a day, that's OK. However, you must not tag more than 50 samples/day.
- Accurate: These data required strong technical English and analytical skills to understand. The tagged data will be further deployed in real research, so accuracy is important.

References:

[1] Tuarob, Suppawong, Sumit Bhatia, Prasenjit Mitra, and C. Lee Giles. "AlgorithmSeer: A System for Extracting and Searching for Algorithms in Scholarly Big Data." IEEE Transactions on Big Data 2, no. 1 (2016): 3-17.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022, 2003.

[3] G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res.(JAIR), 22(1):457–479, 2004.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999

[6] C. Su, Y. Pan, Y. Zhen, Z. Ma, J. Yuan, H. Guo, Z. Yu, C. Ma, and Y. Wu. Prestigerank: A new evaluation method for papers and journals. *Journal of Informetrics*, 5(1):1 – 13, 2011.

[7] S. Tuarob and C. S. Tucker. Fad or here to stay: Predicting product market adoption and longevity using large scale, social media data. In *Proc. ASME 2013 Int. Design Engineering Technical Conf. Computers and Information in Engineering Conf. (IDETC/CIE2013)*, 2013.

[8] D. Walker, H. Xie, K. Yan, and S. Maslov. Ranking scientific publications using a model of network traffi c. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06010, 2007.