Predictions of Citations of a Scholary paper

1st Shreenidhi S
Computer Science Department
College of Engineering Guindy,Anna University
Chennai,India
snidhi1999@gmail.com

2nd Supradeepa Vella
Computer Science Department
College of Engineering Guindy,Anna University
Chennai,India
suprivella88@gmail.com

***Abstract*—Bibliometrics is a statistical analysis of written publications such as books or articles. A bibliographic citation is a reference to a book, article, web page, or other published item. Thus citations are useful for identifying the progress of the particular work and measuring the quality of the research article. The cited papers are downloaded using the crawler. From the downloaded article, identify article relation by analyzing the citation context of the article. So first extract the citation context from the article. Citation context are classifies based on cue phrases of Simon tufel. Next, identify the relation of unlabeled article by word embedding. After labeling all articles identify the perspective behind the citation of the article. In this project, citation relation is identified based on cue phrases of Simon tufel finally article impact is quantified based on the citation network formed from citation analysis.**

***Index Terms*—bibliometrics, citation, word embedding, article**

I. Introduction

Citation context extraction and expansion of citation contexts using various external word sources were attempting to bring more meaningful interpretations into context analysis. Citation context rhetorics have been explored widely in the literature. All these approaches have initiated a semantic analysis using popular similarity metrics like cosine similarity. Starting with cue phrase-based rhetoric analysis to critic analysis, there lies enough scope for text mining incorporated with supervised and unsupervised learning methods to reveal the most out of citation contexts. There is a handful of budding literature on the application of machine learning to bibliometrics analysis.

The network is further embedded with rhetorical citation relations across edges. There might be more than one rhetoric present if the relating nodes are in multiple contexts, which are labeled completely. For enabling rhetoric labeling, we assumed the 12 citation classification categories of Simon Teufel. The 12 citation categories are further classified into positive, negative, and neutral. The rhetoric labels are identified by the matching cue phrases as recommended by Simen Teufel. Upon the absence of cue phrases within the context, the rhetoric is assumed to be neutral. Therefore, the neutral rhetorical category consists of both citations categorized as well as dumb contexts as well. The statistics on rhetoric's of the citation network is presented. The statistics on rhetoric sentiments of citation network is presented. The following section summarises the deliverables obtained from the citation network.

II. Objectives

A. Design a model to analyses the perspective behind the citations.

- To extract the citation context using a text tiling algorithm.
- To classify the context using cue words.
- To classify the context are labeled and unlabeled.
- To appending the labels to the citation network.
- To evaluate the multi labeled citation network using deep learning.
- To evaluate the genetic algorithm based label reduction.

III. Literature Survey

A. Context Extraction

Pert Knoth et al. (2017) proposed system on citation contexts extraction process consists of four stages depicted in we first clean and pre-process the input text. we then process the text line by line classifying each as either a reference line or not. The lines classified as a reference are then passed to a probabilistic parser based on conditional random fields (CRFS) that splits each reference into its constituent fields, such as authors, title, year, and venue. We subsequently use a set of regular expressions to link each reference to all its citations in the processed document extracting all the citation contexts. Finally, we try to link each of the citation reference strings to a unique id of the cited document. Citation context extraction addresses the problem of locating all the links in the body of a paper to each reference and extracting the text surrounding them. There are three main approaches of connecting a citation to its reference we Support:

- The reference is preceded with a number that is used as a citation marker in the body of the document.
- The reference is preceded with an abbreviation created, for example, as a name and year, which is used as a citation marker in the body of the document.
- The citation is linked to a reference using a footnote. We approach the problem of linking the citation to a reference by using a set of regular expressions. These regular expressions were manually created and fine-tuned on a test set. Using a naive base method, the citation context snippet is then formed by a context window of 300 characters to the left and right of the position of the citation.

B. Cue Words

Yulia sibaroni et al. (2016) proposed a system on classified citation-context in a sentence into 12 rhetorical citation categories automatically in three major groups namely weakness, contrast, or comparison. Some rhetorical citations like PBAs, use, modify and most have similar definitions with extends relation definition. Please declare "the author uses cited work as basis or starting point", use declare "the author uses tools/algorithms/data identification", mode declare "author adapts or modifies tools/algorithms/data" and plot declares "this citation is used to motivate works in the current paper". Teufel et al. used a supervised learning approach with the main features is the cue-phrase, while enhancements features are verb tense, voice, and location of a sentence. This approach is interested in develop for identifying extended relations.

C. Labeled and Unlabeled Citation Context

G.S.Mahalakshmi et al. (2018) proposed a system starting from cue phrase-based rhetoric analysis to critic analysis; there lies enough scope for text mining incorporated with supervised and unsupervised learning methods to reveal the most citation contexts .the order of cites appearing in the citing article is also checked for plagiarism with that of the cited article. There is a handful of budding literature on the application of machine learning to bibliometrics analysis. Sincere efforts to employ deep learning techniques to quantify author contributions identify highly cited articles and article topics etc. are being considered in the recent past.

D. Word Embedding

Matthew e. Peters et al. (2018)proposed a full character aware models (Kim et al., 2015) are considerably more parameter efficient than word-based models but more computationally expensive then word embedding based methods when training. During inference, these differences can be largely eliminated by pre-computing embeddings for a large vocabulary and only falling back to the full character-based method for rare words. Overall, for a large English language news benchmark, character aware models have slightly better perplexities than word-based ones, although the differences tend to be small.

E. Deep Learning

R.Siva et al. (2018) proposed deep topic models are topic models constructed over deep-stacked autoencoders. The co-citations extracted for seed article might have overlap with the citation set of seed article and would already be part of the constructed citation graph. therefore, utmost care is observed to remove those in common and the remaining unique co-citations are assumed for further processing the co-citations would not have a direct citation connection with the seed article; additionally, there would be more than one co-cite present in a single context; therefore, yet, there lies an indirect method of computing the same. All the co-cite articles would have some message in common with the seed article, as mentioned by the co-citing research articles. This idea inspired us to obtain the full text of the co-cites excluding references. Since references were not part of the main theme of discussion we chose to ignore them at present. The co-cite full texts are summarized using deep-stacked autoencoder which uses unlabeled data in a complete unsupervised environment to build a compressed representation of input data.

IV. IMPLEMENTATION

The proposed system consists of the following modules:

- Data Collection.
- Context Extraction.
- Citation Context Classification.
- Appending Label to Citation Network.
- To Word Embedding For Unlabelled Data.
- Multi Labeled Citation Network.

V. MODULES

A. Data Collection

The citations of seed paper, "An index to quantify an individual's scientific research output, j.e. Hirsch, 2005" from the bibliometrics domain got downloaded from Google scholar. Later all levels (2005-2016) citation are downloaded from Google scholar both manually and using the crawler. All the citation is downloaded either in pdf or doc format. The pdf files are converted to text and the stop word is removed. For further processing of data to identify the accurate citation relevance. The numbers and symbols are removed after performing the context extraction.

B. Context Extraction

The context where the reference paper is cited is extracted from the citation. The window of 100 words around citations is extracted as referential text. The reference information is used to identify the contexts within the research paper. The system first identifies the paper reference number and search within the document for the presence of this reference number. The context is also extracted using the author's name surrounded by the keywords. The citation index service, which is a representative service of scholarly information services, is gradually showing its limitations. To overcome this, researchers in major foreign countries are pioneering new research areas that use the citation context as a sentence around "in-text citation". These studies extract citation contexts, classify their functions, and use them to try out new citation analysis and citation summaries, or search and visualize citations, and further evaluate the quality of research results.

C. Citation Context Classification

The rhetorical classification is used to identify the different work paths and to identify the positive and negative opinions, emotions, and attitudes expressed in the text. the classification of a cite is identified using cue phrases around the cited area as stated by Teufel. Based on the cue phrases citations are classified into 12 categories. The first category weak is reserved for the weakness of previous research if it is

addressed by the authors. The next four categories describe comparisons or contrast between own and other work. the difference between them concerns whether the contrast is between methods employed or goals (cocogm) or result and in this case of the result, a difference is made between the cited result being worse than the current work (coco) or comparable or better result (cocoro) .as well as considering differences between the current work and other work, citation if they are explicitly compared and contrasted with other work then it is expressed under category coccyx. The next set of categories concerns a positive attitude or opinion expressed towards a citation or a statement that the other work is actively used in the current work. Statements that convey the use of data and methods of the cited work are marked as pure if they are used without any change or marked as pmodi if some adaptations are made. Work which is stated as the explicit starting point or intellectual ancestry is marked with category pbas. If a claim in the literature is used to strengthen the author's argument or vice versa. The similarity of the approach to the cited work is marked as spam and motivation of approach used or problem addressed is marked as pmot. the last category neut bundles truly neutral descriptions of cited work with those cases where the textual evidence for a citation function was not enough to warrant annotation of that category and also used when our scheme did not provide a specific category for the citation.

D. Appending Label to Citation network

The existence of different visualization techniques has been used as a powerful method to enrich visualizing and analyzing the bibliometrics network. A bibliometrics network consists of nodes and links or edge. The node can be an instance of publication, journal authors, or keywords while the edge indicates the relationship between the nodes which can be citation relations or keywords. In this section, we all explain the appending label to the citation network. the citation context is classified based on citation classification categories it denotes as labels .this labels are appending to corresponding article i.e. appending citation network. The network is further embedded with rhetorical citation relations. There might be more than one rhetoric present if the relating nodes are in multiple contexts, which are labeled completely. For enabling rhetoric labeling, we assumed the 12 citation classification categories of Simon Teufel. The 12 citation categories are further classified into positive, negative, and neutral the rhetoric labels are identified by the matching cue phrases as recommended by Simon Teufel. Upon the absence of cue phrases within the context, the rhetoric is assumed to be neutral. Therefore, the neutral rhetorical category consists of both citations categorized as well as dumb contexts as well. The statistics on rhetoric's of the citation network is presented. The statistics on rhetoric sentiments of citation network is presented. The following section summarizes the deliverables obtained from the citation network.

E. Label the unlabeled Data

The word embedding is used to identify the label or unlabeled classification context. The citation context is extracted based on keywords. The keywords are available to extract the context otherwise they cannot extract the context it considers as an unlabeled context. The unlabeled context can be classified using word embedding and extract the context appending the label to the citation network.

F. Multi Labeled Citation network

The article belongs to multiple labels that classify them. To convert the label into a single label using a deep learning algorithm. Finally, append the single label into a citation network in the corresponding article.

VI. CONCLUSION

All the citations of seed articles have been downloaded. From that citation, the citation context has been successfully extracted. Based on simen tufel cue word, it is classified as labeled citation and the unlabeled citation is classified using word embedding. Finally, the citation network is successfully

ACKNOWLEDGMENT

This work would not have been possible without the support of Dr G.S. Mahalakshmi Associate Professor, Department of Computer Science and Engineering,College of Engineering, Guindy, Anna University who guided us throughout the project.

REFERENCES

[1] J. Beel, and G. Bela. "Google Scholar's ranking algorithm: the impact of citation counts (an empirical study)," Proceedings of the Third International Conference on Research C

[2] M. Callaham, L.W. Robert, and W. Ellen. "Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals." Jama 287.21 (2002): 2847-2850.

[3] C. Castillo, D. Debora, and G. Aristides. "Estimating the number of citations using author reputation." String processing and information retrieval, Springer Berlin Heidelberg, 2007.

[4] R. Yan, et al. "Citation count prediction: learning to estimate future citations for literature." Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 20 II.

[5]] R. Yan, et al. "To better stand on the shoulder of giants." Proceedings of the 12th ACMlIEEE-CS joint conference on Digital Libraries. ACM, 2012.

[6] A. Livne, et al. "Predicting citation counts using text and graph mining." Proceedings of the conference 2013 Workshop on Computational Scientometrics: Theory and Applications. 2013.

[7] lE. Hirsch. "An index to quantify an individual's scientific research output." Proceedings of the National academy of Sciences of the United States of America, 102.46 (2005): 16569-16572.

[8] L. Egghe. "Theory and practice of the g-index." Scientometrics 69.1 (2006): 131-152.

[9] L. Egghe. "Theory and practice of the g-index." Scientometrics 69.1 (2006): 131-152.

[10] X. Shi, L. Jure, and A. M. Daniel. "Citing for high impact." Proceedings of the 10th annual joint conference on Digital libraries. ACM, 2010.

[11] P.F. Brown, et al. "The mathematics of statistical machine translation: Parameter estimation." Computational linguistics 19.2 (1993): 263-311.