Important Citations in Paper

By Decoders(Group Number 6)

Mentored by Prof. Purushottam Kar

Shubham Sharma
smsharma@iitk.ac.in

Vikulp Bansal
vikulp@iitk.ac.in

Gurkirat Singh
gurkirat@iitk.ac.in

Dhawal Upadhyay
dhawal@iitk.ac.in

Rahul Gupta
grahul@iitk.ac.in

Abstract

Assigning papers submitted in the conference to their reviewers is one of the most important task in any conference. However this is not a trivial task, large conferences have to assign hundreds of papers to reviewers within a short period of time. The main aim of this project is to use the content and context of the citation within a paper while deciding its best reviewers. We remodel this problem into finding top citations for each paper, which in-turn can be used while deciding its reviewers.

1 Motivation

Conferences usually get a lot of submissions, and the number of reviewers are comparatively fewer. The success of the conference is essentially determined by the quality of papers which get through, and thus it is very important that the researchers reviewing a paper must be selected very carefully.

Existing systems usually ask the reviewers to submit their most “representative" publications, and create a feature map of the reviewers. A feature map is also created for each candidate paper, and using some algorithm for comparing the paper features to reviewer features, the best reviewers are chosen. However, there is a problem with this method. It treats every paper like a newspaper article. It would fail to differentiate if some reviewer’s work has been cited in the paper’s experiment section or the related work section. A citation appearing in the experiment section might correspond to a comparison, which would be highly relevant, as opposed to a citation in related work.

In this project, we have tried to exploit the context of the citation using a variety of features. For a simpler model, we only tackle the problem of finding most important citations for a paper. If the authors of these papers are part of the reviewing committee, they could be given higher weight-age for reviewing the paper. In effect, the output of our model identifies useful citations, which could be used as a feature to another algorithm for identifying reviewers running on top.

2 Related Work

Conferences like SIGGRAPH, KDD and EMNLP have used applications to automate the task of assigning papers to reviewers. Toronto Paper Matching System(1) is one of the systems that compute similarity score between a submitted paper and the reviewer and has been adopted by the leading conferences in both the machine learning and computer vision communities. MyReview is another conference review system that uses machine learning techniques for reviewer matching(7). Conry, Don, Koren, Yehuda, and Ramakrishnan(8) use collaborative filtering methods combined with extra

information about both reviewers and paper to calculate the matching scores. Rodriguez Bollen(9) built co-authorship graphs from the references in the submissions to suggest the initial reviewers.

3 Methodology

3.1 Data

We used Identifying Meaningful Citations's(2) data-set of 465 annoted pairs (cited paper ID, citing paper ID), where all the pairs are taken from the ACL anthology(3). It was categorized as citation types of 'related work', 'comparison', 'using the work' and 'extending the work' with labels 1,2,3,4 respectively. We changed the data-set to only two labels 0 and 1 indicating if the reference is important or not. We considered the citation to be important if it belongs to 'comparison' or 'extending the work' category and not important if it belongs to 'related work' or 'using the work' category. The dataset of papers behind these citations are scraped using BeautifulSoup(4). The text is extracted from PDF files using 'pdftotext2'(5) and then normalized using python scripts. The segmentation of data is done using 'ParsCit'(6) tool.

3.2 Features Used

- **Number of citations**: This feature denotes the number of citations for the cited paper. The more the number of citations, the more meaningful is the cited paper.
- **Citations per section**: This is a collection of features, where each item of the collection denotes the number of citations for the cited paper in that section. Standard sections have been used (like Introduction, Related Work, Experiments etc.), and non-standard sections names have been put under "others" section.

  This feature is essential in deciding the importance of the paper, and would be useful in differentiating between citations in different sections (eg. experiment v/s related work).
- **Number of citations/Total Citations**: It computer the ratio of number of citations of the cited paper to the total number of citations in the citing paper. It distributes the total citations' weight across the cited papers.
- **1/(Number of References)**: This feature gives the inverse of the length of citing paper's reference list. So, the paper with small reference list would give more weight to all its cited papers than the paper with large reference list.
- **Similarity between Abstracts**: This feature computes the tf-idf score to compute the similarity between the abstracts of a citing paper and its cited paper. High similarity means more importance of cited paper for the citing paper.
- **PageRank Score**: It computes the Page Rank score of cited paper. Higher Page Rank score shows more importance of cited paper for the citing paper.
- **Importance of citation in sentence**: Vectorize the words and learn independent classifier to get importance of citation in citing paper. This feature contributes to the new work done in the existing work. It has been discussed in more detail later in this report.

3.3 Problems faced in Implementation

- The xml version rendered by ParsCit often contained some noise in the text. For this reason, an exact matching could not be used to compare titles, and a closest string matching algorithm was used.
- While extracting the number of citations, some times the title in citation did not match the actual paper title, possibly due to multiple spellings, or the author was just lazy to write the complete title. Although most of these cases were handled by closest string matching, at times manual intervention was also needed.
- Section names had slight variations (eg. Analysis v/s Discussion), but these were easily taken care of.

4 Novel Contribution

In addition to the above features, we also created a feature corresponding to the *importance of citation in sentence*. This feature gives the importance of citation derived from the sentence in which it appears in the paper. Since there was no explicit data set or existing relation extraction technique which tells usefulness of subject/object in a sentence, we annotated around 900 sentences ourselves and used a Bag Of Words representation to learn several classifiers. The results observed using these classifiers have been summarized in Table 1.

Classifier	Accuracy(%)	Precision(0)	Precision(1)	Recall(0)	Recall(1)
RBF SVM	77	0.89	0.39	0.82	0.53
Bernoulli Naive Bayes	82	0.84	0.5	0.97	0.16
Multinomial Naive Bayes	82	0.83	0	0.99	0
Gaussian Naive Bayes	71	0.85	0.26	0.79	0.34
Linear SVM	84	0.85	0.75	0.99	0.19
KNN(k=5)	81	0.83	0.33	0.97	0.06

Table 1: Results for feature measuring importance of citation in sentence

5 Experimental Details

5.1 Classifiers Used

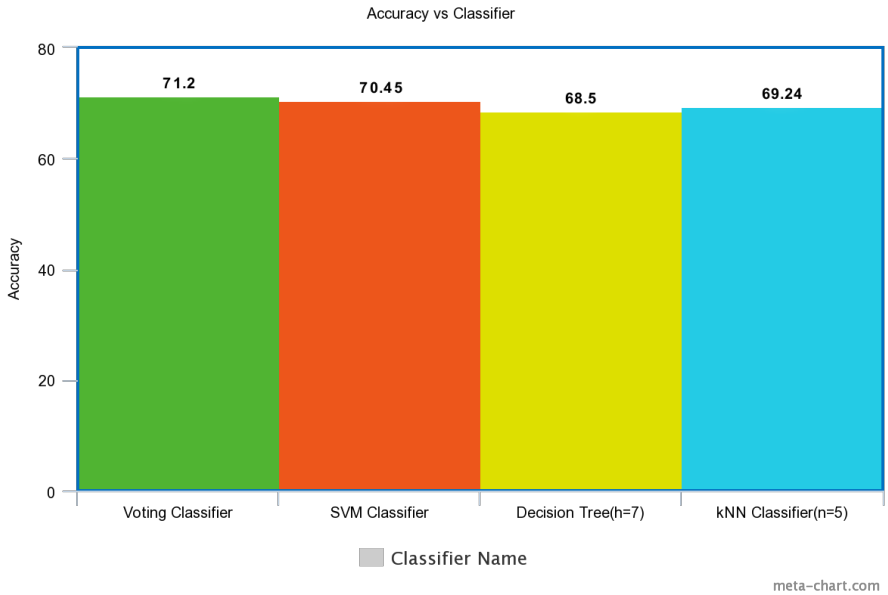
We have independently applied SVM (rbf kernel), Decision Tree, and KNN. The tuning of hyperparameters is described below. We have also used majority voting ensemble classifier with the base classifiers as Decision Tree, KNN, and SVM with rbf kernel.

5.2 Tuning

We used 5-fold cross validation to tune the hyperparmeters. The values for the hyperparameters obtained were -

- Decision Tree: Height = 10
- KNN: number of neighbors = 5

Figure 1 describes the histogram of Accuracy vs Classifier.



6 Future Work

We tried to tackle a smaller problem of finding meaningful citations rather than finding suitable reviewers for a paper. We learned about the working of various supervised classification models. We also studied related work done in this area by going through relevant references. It gave us an understanding of currently prevailing work for paper reviewer matching.

Our current model gives important citations for a paper. These results can be further used to determine suitable reviewers for a paper. The model for its implementation can take these results as a feature and give more weights to authors whose paper's citations got more importance in the citing paper. If these authors are in the review committee, the citing papers can be allotted to authors according to the ranking of weights which were given to authors.

7 Conclusion

Our project attempted to solve the problem of finding most important citations for a paper. We extracted relevant features, which determines cited paper's importance for the citing paper. We have shown various supervised classification approaches using extracted features, with their accuracies. One feature 'Importance of citation in sentence' also gives a new way to tackle the problem of identifying important citations.

References

[1] Laurent Charlin, Richard S. Zemel. *The Toronto Paper Matching System: An automated paper-reviewer assignment system*

[2] Marco Valenzuela, Vu Ha and Oren Etzioni. *Identifying Meaningful Citations*.

[3] http://www.aclweb.org/anthology/

[4] https://pypi.python.org/pypi/beautifulsoup4

[5] https://pypi.python.org/pypi/pdftotext/2.0.0

[6] Isaac G. Councill, C. Lee Giles, Min-Yen Kan. *ParsCit: An open-source CRF reference string parsing package*

[7] Rigaux, Philippe. *An iterative rating method: application to web-based conference management. In SAC, pp. 1682–1687, 2004.*

[8] Conry, Don, Koren, Yehuda, and Ramakrishnan, Naren. Recommender systems for the conference paper assignment problem. Third ACM Conference on Recommender Systems (RecSys-09), pp. 357–360, New York, New York, USA, 2009. ACM. ISBN 978-1-60558-435-5.

[9] Expertise modeling for matching papers with reviewers.