Identifying Citation Sentiment and its Influence while Indexing Scientific Papers

Souvick Ghosh
School of Communication and Information
Rutgers University
souvick.ghosh@rutgers.edu

Chirag Shah
Information School
University of Washington
chirags@uw.edu

Abstract

Sentiment analysis has proven to be a popular research area for analyzing social media texts, newspaper articles, and product reviews. However, sentiment analysis of citation instances is a relatively unexplored area of research. For scientific papers, it is often assumed that the sentiment associated with citation instances is inherently positive. This assumption is due to the hedged nature of sentiment in citations, which is difficult to identify and classify. As a result, most of the existing indexes focus only on the frequency of citation. In this paper, we highlight the importance of considering sentiment of citation while preparing ranking indexes for scientific literature. We perform automatic sentiment classification of citation instances on the ACL Anthology collection of papers. Next, we use the sentiment score in addition to the frequency of citation to build a ranking index for this collection of scientific papers. By using various baselines, we highlight the impact of our index on the ACL Anthology collection of papers. Our research contributes toward building more sentiment sensitive ranking index which better underlines the influence and usefulness of research papers.

1. Introduction

Our work toward developing a sentiment-sensitive ranking index for scientific papers can be situated at the intersection of bibliometrics, real-world citation networks, and sentiment analysis. A graphical representation of a hypothetical citation network has been presented in Figure 1. Each node of the graph represents a scientific paper in the collection. In scientific papers, we could find mentions of other papers. These mentions, called citations, reflect the view of the author (of the source paper) towards the target paper. We can visualize these instances of citation as directed edges which originate from the source or citing paper and point to the target or cited the paper. Previous studies [1, 2, 3, 4] have revealed that citation networks exhibit the properties of the small-world network with high clustering coefficient and small degrees of separation. This highlights that a lot of citations are observed within a closed community and as such the criticisms are often expressed in polite terms.

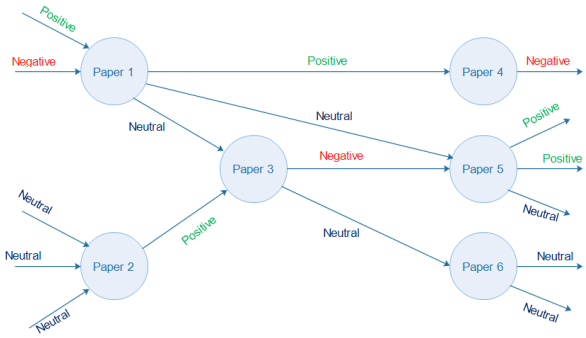

Figure 1: Example of Citation Network.

The lifecycle of most research projects begins with a concept or an idea and ends with a publication in a conference, journal, or any other suitable venue. If one explores the collection of scientific papers in a given field or research area, one could identify a directed network between the papers and the authors as they cite each other in their respective works. Investigating this network of scientific citations has been the focus of research in computer and information science. Looking at the ensemble of papers, we could identify the relative importance of the papers, the authors, and the ideas expressed in the papers. We could also identify how the different entities – papers, authors, and ideas – are connected to each other in the network of citations. It would allow the researchers to identify the most influential papers in the network and their degree of influence on the other papers [1]. In some cases, the absence of citations could serve as a significant clue

URI: https://hdl.handle.net/10125/64049
978-0-9981331-3-3

in determining areas of future research (unexplored domains or novel ideas) [5].

The network of citations was the first step towards determining the influence or impact of the scientific publications on the scientific community. Various bibliometric indicators have been designed which focus on the total number of papers, the total number of citations, or on the number of papers published in high-quality venues. Some of the major drawbacks of such an approach are that the number of citations does not reflect the influence of these papers on the research area. The frequency-based indicators can be biased towards authors with few highly cited papers. Also, such ranking systems fail to recognize highly productive young authors.

In a citation network, each of the citation instances is an ideal candidate for analyzing sentiments. This would be similar to assigning a sentiment based edge weight to each of the directed edges in the citation network. However, sentiment analysis of citation instances is a relatively unexplored area of research. Although most of the real world citations are objective in nature [6], i.e., they present facts and findings without expressing any opinion, yet there is a common assumption that most research papers are cited positively [6]. Athar (2014) [6] mentions the "sociological aspect of citing," which prevents researchers from strongly criticizing their peers. The criticisms are often expressed in polite, contrasting terms. The negative citations, if present, are implicit or hidden, which poses a major challenge for detection. The examples of the three different types of sentiments in citations have been shown in Figure 2. Sentiment analysis is genre-specific in nature, and scientific literature differs vastly from the other forms of text. Positive and negative citations are not necessarily "good" or "bad", instead, it involves identifying the polarity of the opinion of the citing paper towards the cited paper for a particular citation instance (praise or criticism of a specific aspect). When multiple instances of citation are observed between two papers, it is a common practice to identify the sentiment of each instance separately instead of trying to assess the overall polarity. For example, Paper A may positively mention different aspects (say algorithms used and experimental design) of Paper B but may be critical of some other aspect (say evaluation techniques used). The overall sentiment could be calculated as a function of individual sentiments.

In our paper, we argue that it is unfair to evaluate and rank scientists rather than the papers themselves. Ordering a collection of paper would serve a myriad of purposes like providing high visibility for top quality papers and identifying gaps in research. Most of the existing bibliometric evaluation schemes [7, 8, 9, 10] focus primarily on the quantitative aspect of citations. It is often noticed that some of the good papers are never cited while some of the poor quality papers receive a lot of citations which are mostly negative in opinion (for the purpose of criticism). Therefore, for a fair and accurate analysis of the influence of the paper, we need to consider the qualitative aspect of citations, that is, the tone and the polarity of sentiment expressed by the citing paper.

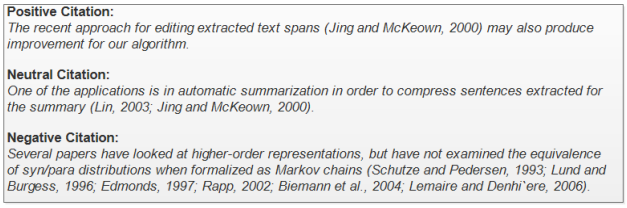

Figure 2: Example of Different Sentiments in Citation.

For this research we have the following research objectives:

1. Determine the sentiment for each citation instance in a comprehensive collection of research papers.
2. Formulate a new ranking index that considers the quality as well as the quantity of citations.
3. Evaluate the impact of the proposed index on the ranking of research papers in the collection.

The following three research questions guide the overall direction and objective of this research:

RQ1: What is the reasonable accuracy with which we can measure the sentiment associated with scientific citations?

RQ2: How can we use the sentiment of citation to build a new ranking index for bibliometric purposes?

RQ3: What are the impacts of qualitative measures on the ranking of research papers, which has been dependent on only quantitative indicators?

The rest of the paper is organized as follows: We review the related work in the next section. In the following sections, we describe the preparation of the corpus and the lexical resources, our approach for automatically identifying the sentiment in citations, and the different ranking indexes and their influence. In the last section, we conclude the paper and suggest a direction for future work.

2. Related Work

Although a lot of work has been done in areas of Sentiment Analysis, Scientometrics, and Network Theory separately, there are only a few studies which consider the sentiment of citation while assessing the

impact of scientific papers. In the following paragraphs, we discuss some relevant research in the three domains which guide the design and approaches used in this study.

Sentiment Analysis is an interesting research area due to the increasing popularity of various online platforms (not limited to social media), the reduced cost of collecting and storing information online, and the increased efficiency with which that data could be processed. The availability of a large amount of textual data and the development of sophisticated text processing techniques has facilitated the analysis of sentiment in a wide variety of documents. Such analysis is not limited only to movie or product reviews but could also be extended to stance detection in politics and social media, and citation networks. Majority of sentiment analysis tasks follow a two-step model [11, 12] where the first step is to detect if the given sentence or document has any subjectivity associated with it. Once this binary classification has been made (Subjective or Objective) [13], the next step is to classify the polarity of all the subjective instances. However, the task of assigning sentiment to citation instances gets more complex as the expressed sentiment is often implicit and hard to identify using off-the-shelf classifiers. Scientific community use domain-specific lexicons and terminology while citing papers and as such, it is necessary to identify the literature specific lexicons for analyzing sentiment in scientific papers.

Researches in scientometrics have covered some interesting but diverse applications: authors have tried to understand the citation behavior by determining the interrelationship between citation type, utility, and location [14]; used co-citation analysis to connect two literature and solve a problem [15]; analyzed the influence of the editor on citation patterns [16]; and explored the cost of collaboration for authors of retracted papers [17].

Citation network analysis, on the other hand, provides an insight into the patterns of citation observed in a research community. Newman (2003) [2] performed an empirical analysis of networked systems like the Internet, and social networks. Some other works [3, 4] investigated the degrees of separation in collaboration networks of scientific journals. Travers and Milgram's (1967) [18] experiments explained the small world problem, where every individual is linked to others in a closely-knit societal structure, with six degrees of separation. Elmacioglu and Lee (2005) [3] showed that the Digital Bibliographic Library Browser (DBLP) network resembles a small-world network where the average distance of all the scholars is approximately six. Few scholars publish a large number of papers while the majority publishes little. Also, with "publish or perish" outlook in academia and research, scholars feel the need to collaborate, thus displaying increased clustering coefficients in the collaboration network. Similarly, Nascimento et al. (2003) [4] showed that the co-authorship graph of SIGMOD exhibits the properties of small-world network with a clustering coefficient of 0.69 and characteristic path length of 5.65. Rahm and Thor (2005) [19] analyzed the citation frequencies of conferences and journals papers over ten years and determined the most cited authors, authors, their institutions, and countries. Leicht et al. (2007) [20] investigated large-scale citation networks which evolved over time. They used three different approaches – expectation-maximization, modularity optimization, and eigenvector centrality – to demonstrate the structural divisions in the network. They hypothesized that by highlighting the qualitative changes in the citation patterns, we could picture the community structures present in such a network. Shi et al. (2009) [21] investigated how the proximity in the subject area between the citing and the cited paper influences the impact of the citing publication. All these research works highlight the fact that citation networks are often highly clustered and dense networks where the authors cite their peers, mentors, and collaborators. Due to the small and closed nature of research communities, it becomes necessary to hedge the criticisms.

Existing bibilimetric measures considers all citations positively even though some of them are criticisms. Therefore, the impact or usefulness of a paper should not be dependent only on the number of times it is cited but also on how it is cited. In other words, the opinion of the citing authors can be leveraged to assess the influence of any paper. Spiegel-Rosing (1977) [22] was one of the first researchers who pointed out the "lack of any content analysis of citations – especially of the evaluative component of citations (critical/appraising) – has been a major point of criticism in the use of citations as an indicator of the quality of research." (Spiegel-Rosing, 1977, p.101) [22]. Small (1978) [23] proposed that the citations should be interpreted as concept symbols as they reflect the previous knowledge on which the author has constructed the current paper. Over the last decade, we have observed a growing trend toward automatically analyzing the sentiment associated with citations [24, 25, 26, 27]. Bonzi (1982) [28] investigated how the different characteristics of the citing and cited paper help in assessing the relatedness between the papers. Some of the characteristics investigated were the source of both the papers, the number of citations, self-citations, the year of publication, the type and length of the citing article, the placement of the citation,

and the sex of the authors. Teufel et al. (2006) [26] investigated the discourse of citation and suggested that the automatic recognition and classification of such functions could help in replicating human annotation and increase performance in large scale environments. In their work, the authors used supervised learning with linguistic features for classifying the citation functions.

Abu-Jbara et al. (2013) [29] criticized the existing bibliometric measures as they do not differentiate between positive and negative citations. They used natural language processing techniques to assess the sentiment of the paper. By using supervised machine learning techniques, they identified the occurrence, the purpose, and the polarity of sentiment in the citation instances. Athar (2011) [30] explored the usefulness of n-grams, specialized lexicons for scientific literature, dependency relations, sentence splitting, and negation features for sentiment analysis. The author suggested that n-grams and dependency-based features performed best in the classification task. Athar and Teufel [24, 25] addressed the problem of context-enhanced detection of citation sentiment.

In our work, we have combined the techniques in natural language processing, sentiment analysis, and network theory to develop a sentiment enhanced ranking algorithm for scientific papers. We have used a combination of natural language features for generating sentiment scores for each citation instance. We constructed a directed network graph using each paper as a node and each citation instance as an edge. The edge weights were combined to evaluate the overall influence of the scientific paper.

3. Corpus Preparation

The sentence which contains the citation has been referred to as citation sentence, the source and target papers are the citing and cited paper respectively(Figure 3). While some of the recent works in citation sentiment analysis have adopted a context based approach (where sentences before and after the citation sentence is used for analysis), using a context enhanced approach calls for resolving the overlapping scopes of citation. In this work, we have concentrated on only the citation sentence for analyzing the sentiment. We have used two corpora for our work. The first corpus has been obtained from work done by Athar (2011) [30]. The dataset contains a total of 8736 citation instances, with each instance containing a single sentence. An example of citation instance has been shown in Figure 3. The dataset contains the following information: Citation sentence, citing paper identifier (or the source), the cited paper (or the target), and the sentiment of citation. We used this corpus for training and testing our sentiment classification model. Keeping it consistent with the general practice in machine learning, we used around 80% of the dataset (6736 instances) for training and the rest (2000 instances) for testing purposes. After evaluating the classifier for accuracy, we used the entire dataset for retraining the classifier.

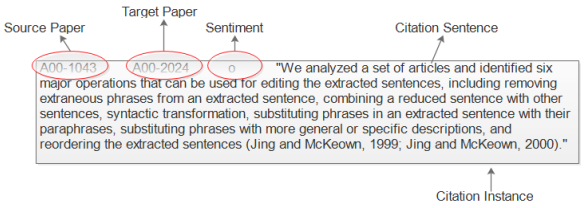

Figure 3: An Example of Citation Instance.

We used a second corpus for exploring the sentiment based ranking algorithm. This larger corpus, consisting of more than 77000 citation sentences, was prepared from the ACL Anthology Network (AAN). The Association for Computational Linguistics (ACL) is "*the premier international scientific and professional society for people working on computational problems involving human language; a field often referred to as either computational linguistics or natural language processing (NLP)*[1]." ACL Anthology Network[2] is the collection of all papers in the venues under ACL. In our work, we automatically detect the various citation patterns in each paper using regular expressions and create a collection in structured file (in XML). This file is further processed to separate each citation instance along with the source and target paper identifiers. As this collection was not annotated, we used our sentiment classifier (accuracy reported in Section 5.3) to automatically assign sentiments to each instance. After annotation, the dataset consisted of 52491 neutral instances, 10064 positive instances, and 10326 negative instances. This classifier is specifically meant to detect negative polarity in scientific papers and the high number of negative instances detected attest to the high recall of our classification model.

Table 1: Distribution of Instances in the Corpora.

Corpus	Positive	Neutral	Negative
Corpus 1 (Cambridge)	809	7322	272
Corpus 2 (AAN)	10064	52491	10326

[1]https://www.aclweb.org/portal/what-is-cl
[2]http://clair.eecs.umich.edu/aan/index.php

4. Creation of Lexical Resources

The training data was used for creating a set of specialized sentiment words used in scientific literature. We used a semi-supervised approach (bootstrapping) for identifying the set of subjective words in the citation instances. We started with a list of seed words and identified subjective sentences using a word matching algorithm. For each subjective sentence, we manually identified the science-specific lexicons and added them to our list of seed words. The list of most frequent unigrams, bigrams, and trigrams for each polarity are presented in Table 2, 3 and 4.

Table 2: Most Frequent Unigrams with Positive Polarity.

More	Improvement	Outperform	Popular
Most	Important	Correlate	Efficient
Improve	High	Higher	Successful
Best	Effective	Major	Overcome
Well	Accurate	Significant	Consistent
Better	Development	Highly	Sophisticated
Simple	Useful	Robust	Benefit
Good	Successfully	Considerable	Simpler

Table 3: Most Frequent Positive n-grams (n>1).

improvement in	success of
more efficient	good performance
very successful	can improve
most successful	most notable
well known	good result
more accurate	best score
development of	most important
widely used	high quality
achieve impressive	quite accurate
effective at	increase over
state of the art	improve performance of
very high frequency	most widely used

Table 4: Most Frequent Negative Words.

However	Unlike	Worse
Unrealistic	While	Restrict
Unfortunately	Insufficient	Although
Lack	Complicated	Inability
Low	Poor	Daunting
Lack of	Without	Unexplored
Degrade	Not well	Difficult
Little	Burden	Not able to

5. Automatic Sentiment Classification

5.1. Preprocessing

The annotated corpus (Corpus 1) was used for developing the automatic sentiment classifier. The distribution of instances of different polarity is shown in Table 5. As the training data was biased towards the neutral class, we applied SMOTE technique [31] to deal with the imbalances in the dataset. SMOTE oversamples the minority class in the dataset by generating new, synthetic examples in the feature space using nearest neighbor approach. This is an improvement over simple replication of data points which often affects the accuracy of the classifier. After SMOTE, our training data contained a balanced representation of all the three sentiment classes. We trained our classification model using this balanced dataset and evaluated the accuracy of the model using realistic, unbalanced test data (accuracy reported in Section 6.3).

Table 5: Instances of Each Polarity in Corpus 1.

Dataset	Positive	Neutral	Negative
Training	619	5644	206
Training (with SMOTE)	4952	5644	3296
Test	190	1678	66

5.2. Features for Sentiment Classification

In our work, our focus was on developing a simple, yet accurate sentiment classifier which could be easily scaled for a large real-world dataset. The features selected were motivated by the literature [30, 26, 28, 27] and have been explained in the following subsections.

1. **Automatic Sentiment Score:** For each citation instance, the automatic sentiment score was calculated by splitting the citation sentence into a bag of words. After normalizing all the words to their stems, we have calculated the sentiment score for each of them using SentiWordNet. [3] The sentiment score of each sentence is obtained by summing up the scores of the individual words in the sentence. e.g.: *Dasgupta and Ng (2007) improves over (Creutz, 2003) by suggesting a simpler approach. (Citing paper id 'W09-0805', cited paper id 'N07-1020', sentiment score: 43.0)*
2. **N-grams with positive polarity:** For each citation sentence, we find out the number of words which matches the list of positively polar n-grams specific in scientific literature (described in Tables 2 and 3)
3. **N-grams with negative polarity:** We find the number of matches between the words in the citation sentence and the list of negative words pertaining to scientific literature (described in Table 4)
4. **Presence of specific part-of-speech tags:** Each of the citation sentences is analyzed using a part-of-speech tagger. The presence of specific part-of-speech tags like JJ, JJR, JJS, JJT (various forms of adjective), RB, RBR, RBT, RN, RT (forms of adverbs) and FW (foreign

[3] http://sentiwordnet.isti.cnr.it/

words) are indicators of subjectivity [27]. Similarly, the occurrence of adverbs followed by an adjective (e.g. RB_JJ) is a strong indicator of sentiment polarity in the sentence. e.g.: *simpler/ JJR, well/RB, etc.*

5. **Presence of specific dependency tags:** After constructing the dependency tree for each citation instance, we searched for the different dependency tags which are indicators of subjectivity in citation sentence. Examples of such tags are *advmod* (adverb modifier), *acomp* (adverbial complement) and *amod* (adjectival modifier). e.g.: *simpler approach, well known*, etc [27]. Here, *amod (approach, simpler)* and *advmod (known, well)* captures the polarity of citation. Similarly, acomp functions like an object of the verb and amod are any adjectival phrase that modifies the meaning of the NP. These relations are most frequent in sentences where polar sentiments are present.

6. **Presence of self-citation:** We check the citation sentences where the authors site their own research. This is a straightforward task as we need to verify if the source and the target papers are the same.

7. **Presence of other sentiment words:** To identify subjectivity in citation text, we used a publicly available list of polar words [32]. This collection contains words which denote positive and negative sentiment on the Web. In addition, we used another dictionary of sentiment-bearing words, called Vender Sentiment, which contains all the words and their most likely associated sentiment scores. We divided these lists into two collections each – one for positive sentiment and the other for negative sentiment – and introduced four features. Each feature is a count of the number of matches with each list.

5.3. Classification Results

For classification of the instances into three sentiment classes, we used the machine learning software WEKA[4]. Using the features described in the previous section, we have trained our model using different classifiers, out of which Dagging performed the best. Dagging is a meta-classifier which divides the data into a number of mutually exclusive stratified folds. Each set of data is classified using some base classifier. The final prediction is made using majority voting. The results of the classification are presented in Table 6. The overall accuracy of the classifier was **80.61**%. The confusion matrix is presented in Table 7.

[4]http://www.cs.waikato.ac.nz/ml/weka/

Table 6: Detailed accuracy by class.

	Precision	Recall	F-measure	ROC Area
Positive	0.37	0.6	0.46	0.76
Neutral	0.94	0.85	0.89	0.78
Negative	0.17	0.27	0.21	0.73

Table 7: Confusion matrix for classification.

	Positive	Neutral	Negative
Positive	114	67	9
Neutral	172	1427	79
Negative	22	26	18

6. Ranking Indexes

After obtaining the classification model, the sentiment of each citation instance in the larger corpus (Corpus 2) was annotated automatically. We formulated four different approaches to ranking the papers in the collection which are explained in the following sections.

6.1. Ranking Index 1 (R1-Index)

The R1-index for a node n is calculated by finding the total number of incoming nodes for a given node. This is a simple indexing approach which assumes that the importance of a paper can be assessed by counting the number of citations it receives. This is strictly a quantitative measure of the citations without any provisions for measuring quality.

6.2. Ranking Index 2 (R2-Index)

The R2-index is a modification of the R1-index. It is based on the belief that in order to measure the relative importance of any research paper, we must consider both the quantitative and qualitative aspects of the citation. To calculate R2-index for any node n, we assign sentiment score (from the classification step) to each directed edge (which represents a citation instance) $m \rightarrow n$ where m is the source and n is the target node.

6.3. Ranking Index 3 (R3-Index)

R3-index is a modification over R1-index as it considers the link structure to determine the importance of the cited paper. The working of this index is similar to Google's PageRank algorithm [33] where the number and quality of the links help in determining the importance of the page . It is assumed that as the importance of a page (paper in case of citation networks) increases, the number of links (directed edges in our case) it receives also increases. For any directed edge $m \rightarrow$n, the paper m transfers some of its score to the

cited paper n. If the citing paper cites a large number of papers, then the score transferred to each of the cited paper decreases proportionately. One of the possible shortcomings of this index is that it does not consider the sentiment of the citation, that is, it does not consider if the citation is a general statement, praise, or criticism.

6.4. Ranking Index 4 (R4-Index)

R4-index is a major step towards the sentiment-sensitive ranking of scientific papers. It is similar to R3-index as it uses the same approach in estimating the importance of the research paper through link analysis. However, unlike R3-index, it considers the associated sentiment while determining the edge weight. For any directed edge $m \rightarrow$n, the score transferred from m to n is multiplied by the sentiment score associated with that citation instance. The R4-index for node n is the total of all the incoming scores from the citing papers.

6.5. Evaluation of the Ranking Indexes

The evaluation of the ranking indexes is similar to comparing two or more ranked lists with each other. The process of comparing ranked lists has been much explored in the literature for recommendation systems. As each of the indexing approaches generates a different ordering of papers in the collection, it is essential that we investigate if one ranked list differs significantly from the other. To quantify the degree of similarity or the difference between two ranked lists, we have followed two different approaches:
1. Ranked Correlation; and
2. Set Based Measure.

6.5.1. Ranked Correlation We have used Kendall rank correlation coefficient for measuring the probability that two papers belong in the same order in the two ranked lists. For example, if paper A appears before paper B in one list, then we calculate the probability that paper A precedes paper B in the second ranked list. This approach allows us to evaluate how the relative ordering of the scientific papers changes between two ranked lists. In other words, it helps us in evaluating the similarity between any two indexes. Kendall ranked correlation for the four indexes R1, R2, R3 and R4 are presented in Table 8.

In Table 8, the diagonal elements are all 1s, which shows that any index is perfectly correlated to itself (as there are no changes in the relative ordering of papers). However, if we look at the last column of the table, we can find that the correlation of R4-index with R1 and R2 indexes are quite low. As R4-index differs from R3-index only in the inclusion of sentiment, a correlation of 0.714 can be observed. However, for ranked lists, a correlation of 0.714 highlights that some papers have changed their ordering between the two indexes.

Table 8: Correlation Between the Indexes.

	R1-index	R2-index	R3-index	R4-index
R1-index	1.000	0.790	0.571	0.394
R2-index	0.790	1.000	0.484	0.493
R3-index	0.571	0.484	1.000	0.714
R4-index	0.394	0.493	0.714	1.000

*All correlation values were significant (p<0.001), n = 15260

While Kendall rank correlation coefficient can suffer if the there are missing items in one of the lists, all our ranked lists contain the same set of papers. Another shortcoming of the above correlation measure is that it does not differentiate between the ranks of the papers as long as the relative ordering remains the same. However, in the case of ranked lists, it is often considered that top-ranked items are more important than the lower-ranked items. Therefore, the correlation score should be lower if switching of paper order occurs at the top of the list than at the bottom. We have mitigated this problem through our next evaluation measure - the set based approach.

6.5.2. Set Based Measure The set based measure considers the two ranked lists as a bag of items and aims to find the number of intersections between the two bags. By measuring the set intersections at different depth of the ranked lists, it is possible to quantify the degree of similarity. The different scores which can be calculated using the set based approach are as follows:
1. **Overlap at level-k:** The number of matches found in top-k papers between the two ranked lists.
2. **Average Overlap Score:** For any level-k, we obtain the overlap at level-k and express it as a fraction of the maximum number of possible matches. The advantage of using this score is that it is bounded between 0 and 1.
3. **Rank Biased Overlap:** For any value of k, we calculate the average overlap score for depths 1 to k. We add the average overlap scores and divide it by k to obtain the rank biased overlap [34]. Using this approach ensures that observing a common paper at higher rank contributes to all the lower ranked intersections; therefore, this score is sensitive to the movement of higher ranked papers.

The Rank Biased Overlap scores are calculated in Table 9. The number of pair wise matches is given in parenthesis. The scores and the number of matches

Table 9: Rank Biased Overlap Scores for the 4 indexes.

		R1-index	R2-index	R3-index	R4-index
R1-index	Top-10	1.00 (10)	0.85 (9)	0.11 (2)	0.21 (2)
	Top-100	1.00 (100)	0.91 (89)	0.32 (38)	0.33 (36)
	Top-500	1.00 (500)	0.89 (432)	0.41 (224)	0.41 (216)
	Top-1000	1.00 (1000)	0.88 (468)	0.45 (525)	0.44 (510)
R2-index	Top-10	0.85 (9)	1.00 (10)	0.19 (2)	0.23 (2)
	Top-100	0.91 (89)	1.00 (100)	0.34 (38)	0.35 (38)
	Top-500	0.89 (432)	1.00 (500)	0.41 (221)	0.42 (225)
	Top-1000	0.88 (468)	1.00 (1000)	0.44 (513)	0.45 (516)
R3-index	Top-10	0.11 (2)	0.19 (2)	1.00 (10)	0.79 (7)
	Top-100	0.32 (38)	0.34 (38)	1.00 (100)	0.72 (72)
	Top-500	0.41 (224)	0.41 (221)	1.00 (500)	0.73 (380)
	Top-1000	0.45 (525)	0.44 (513)	1.00 (1000)	0.74 (747)
R4-index	Top-10	0.21 (2)	0.23 (2)	0.79 (7)	1.00 (10)
	Top-100	0.33 (36)	0.35 (38)	0.72 (72)	1.00 (100)
	Top-500	0.41 (216)	0.42 (225)	0.73 (380)	1.00 (500)
	Top-1000	0.44 (510)	0.45 (516)	0.74 (747)	1.00 (1000)

are highlighted for R4-index. It shows that although are R3 and R4 are similar with a score of 0.74 for top 1000 documents, yet there are only 747 papers which maintain their positions in the two ranked lists. For top 10 papers, three papers have changed their positions between the two lists. Overall, including the sentiment changes the ranking by at least 25% between R3 and R4 indexes. When comparing the R4-index to others, the similarity drops to below 50%. Table 10 gives the overlap scores at level-k for different values of k. It is interesting to note that out of 6 pairs of overlap scores, only the overlap scores between R1 and R2 indexes are monotonously decreasing as the value of k increases from 1000 to 5000. This shows that there is a greater degree of overlap of top-ranked papers between the two lists.

6.6. Influence of citation sentiment in ranking

In this section, we investigate the influence of citation sentiment on ranking. We have selected a particular paper (Paper ID: P03-1021) which has a ranking of 2, 5, 27, and 22 in R1, R2, R3, and R4 indexes respectively. The paper has a total of 574 citations. Our sentiment classifier model annotated 72 of them as involving negative sentiment, 39 as positive, and the rest as objective (or neutral).

We have picked two examples where P03-1021 is cited negatively.
Example 1:
In equation 8, there are two types of parameters: parameters introduced by the gain function and the model cost, and system weights introduced by the mixture model; because equation 8 is not linear function when all parameters are taken into account, Mert algorithm (Och, 2003) cannot be directly applied to optimize them at the same time.
Example 2:
The ubiquitous minimum error rate training (Mert) approach optimizes Viterbi predictions, but does not explicitly boost the aggregated posterior probability of desirable n-grams (Och, 2003).

We have also selected two instances where P03-1021 has been cited positively:
Example 1:
The most popular algorithm for this weight optimization [sic] is the line-search based MERT (Och, 2003).
Example 2:
Those weighting coefficients can be learned from the development set via the well-known Minimum Error Rate Training approach (Schluter and Ney 2001; Och 2003) (commonly abbreviated as MERT).

The paper rank changes from 2 in R1-index to 5 in R2-index. This movement is only due to the presence of negative citations (or criticisms). R3-index considers link structure for ranking while R4-index is sensitive to both sentiment and link structure. As the link structure is dependent on the citing papers, it is hard to identify the exact instances which lowered the ranking of P03-1021 in both of these indexes. However, as both R3 and R4 consider the quality of the citations (as reflected by the links and sentiments), it could be speculated that including qualitative measures like sentiment influenced the ranking of P03-1021. As most of the existing indexes focus on quantitative measures (the number of citations), our indexes (R4-index in particular) offer an alternate approach to rank scientific papers.

7. Conclusions and Future Work

In this paper, we explored the influence of sentiment on citations and how we can use it for ranking scientific papers. The contributions of this paper can be divided into three broad categories. The first research objective

Table 10: Overlap at Level-k.

	k	R1-index	R2-index	R3-index	R4-index
R1-index	1000	1.00 (1000)	0.868 (868)	0.525 (525)	0.510 (510)
	2000	1.00 (2000)	0.865 (1729)	0.640 (1280)	0.591 (1182)
	3000	1.00 (3000)	0.856 (2568)	0.677 (2033)	0.620 (1860)
	4000	1.00 (4000)	0.850 (3400)	0.703 (2812)	0.650 (2599)
	5000	1.00 (5000)	0.849 (4243)	0.746 (3730)	0.676 (3379)
R2-index	1000	0.868 (868)	1.00 (1000)	0.513 (513)	0.516 (516)
	2000	0.865 (1729)	1.00 (2000)	0.611 (1222)	0.601 (1202)
	3000	0.856 (2568)	1.00 (3000)	0.657 (1970)	0.643 (1929)
	4000	0.850 (3400)	1.00 (4000)	0.675 (2700)	0.680 (2721)
	5000	0.849 (4243)	1.00 (5000)	0.704 (3521)	0.708 (3541)
R3-index	1000	0.525 (525)	0.513 (513)	1.00 (1000)	0.747 (747)
	2000	0.640 (1280)	0.611 (1222)	1.00 (2000)	0.783 (1566)
	3000	0.677 (2033)	0.657 (1970)	1.00 (3000)	0.795 (2385)
	4000	0.703 (2812)	0.675 (2700)	1.00 (4000)	0.807 (3229)
	5000	0.746 (3730)	0.704 (3521)	1.00 (5000)	0.822 (4109)
R4-index	1000	0.510 (510)	0.516 (516)	0.747 (747)	1.00 (1000)
	2000	0.591 (1182)	0.601 (1202)	0.783 (1566)	1.00 (2000)
	3000	0.620 (1860)	0.643 (1929)	0.795 (2385)	1.00 (3000)
	4000	0.650 (2599)	0.680 (2721)	0.807 (3229)	1.00 (4000)
	5000	0.676 (3379)	0.708 (3541)	0.822 (4109)	1.00 (5000)

was to accurately measure the sentiment associated with scientific citations. We developed a supervised machine learning classifier to automatically identify the sentiment in each citation sentence. The classifier makes use of various natural language techniques and lexicons specific to scientific papers. It has an accuracy of 80.61% with high recall for positive and negative sentiments in citations. This proves that our classifier successfully identifies implicit criticisms and negative opinions common in scientific papers.

The second objective was to use the citation sentiment to build a new ranking index. To this end, we used our classifier to assign sentiment scores to every citation instance in a large dataset obtained from ACL Anthology Network. The ACL citation network consists of 15,260 scientific papers and more than 77,000 citations. We assigned sentiment score specific to each citation instance and used it for calculating the overall score. The overall citation score from citing to cited paper was obtained by either summing all the individual scores (for R2) or by using edge weight and PageRank (for R4).

Our last objective was to assess the impacts of qualitative measures on the ranking of research papers. We used different measures like frequency, sentiment and link analysis to develop four different indexes - R1 and R3 considers only frequency and only link structure respectively, R2 considers frequency and link structure while R4 considers all three factors (frequency, sentiment, and the link structure). While most of the existing indexes are focused towards quantitative assessment, the R4-index is a step towards sentiment enhanced ranking of scientific papers. We demonstrated that inclusion of sentiment and link structure leads to at least 25% difference between two ranking indexes. The change of rank for papers show the huge influence of the sentiment associated with citations. When a highly influential paper is cited but the citing paper points out the limitations, it indicates a negative sentiment. It may be argued that including such instances to boost the ranking of the paper is unfair to newer or less influential papers. Frequency based ranking have often been criticized for favoring more influential papers which leads to more citations cyclically. A sentiment-based approach might mitigate such outcomes. It must be noted that while this show the influence of sentiment on the ranking, more research is required to provide insights on the usefulness of sentiment-enhanced ranking to the end users.

In future, we plan on conducting a crowdsourced study to obtain human evaluation of the sentiments and the revised ranked lists. It would be interesting to assess how negative citation influences the perception of readers towards the negatively cited paper. We would also like to perform a detailed analysis of the correlation patterns between different ranking approaches. This would help us to identify the reasons which influence the differential ordering of the papers under different ranking schemes.

References

[1] D. R. Radev, M. T. Joseph, B. Gibson, and P. Muthukrishnan, "A bibliometric and network analysis of the field of computational linguistics," *Journal of the American Society for Information Science and Technology*, vol. 1001, pp. 48109–1092, 2009.

[2] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[3] E. Elmacioglu and D. Lee, "On six degrees of separation in dblp-db and more," *ACM SIGMOD Record*, vol. 34, no. 2, pp. 33–40, 2005.

[4] M. A. Nascimento, J. Sander, and J. Pound, "Analysis of sigmod's co-authorship graph," *ACM Sigmod record*, vol. 32, no. 3, pp. 8–10, 2003.

[5] D. R. Swanson, "Response: Absence of citations can be valuable clue," *Journal of the American Society for Information Science*, vol. 40, no. 3, pp. 152–152, 1989.

[6] A. Athar, "Sentiment analysis of scientific citations," tech. rep., University of Cambridge, Computer Laboratory, 2014.

[7] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National academy of Sciences of the United States of America*, pp. 16569–16572, 2005.

[8] J. E. Hirsch, "An index to quantify an individualâĂŹs scientific research output that takes into account the effect of multiple coauthorship," *Scientometrics*, vol. 85, no. 3, pp. 741–754, 2010.

[9] L. Egghe, "Theory and practise of the g-index," *Scientometrics*, vol. 69, no. 1, pp. 131–152, 2006.

[10] E. Garfield, "The impact factor," *Current contents*, vol. 25, pp. 3–4, 1994.

[11] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86, Association for Computational Linguistics, 2002.

[12] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, p. 271, Association for Computational Linguistics, 2004.

[13] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara, "Development and use of a gold-standard data set for subjectivity classifications," in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 246–253, Association for Computational Linguistics, 1999.

[14] V. Cano, "Citation behavior: Classification, utility, and location," *Journal of the American Society for Information Science*, vol. 40, no. 4, pp. 284–290, 1989.

[15] H. Small and E. Garfield, "Analysis of scientific literature to assist in problem solving," *Journal of the American Society for Information Science*, vol. 40, no. 3, pp. 152–152, 1989.

[16] M. Sievert and M. Haughawout, "An editor's influence on citation patterns: A case study of elementary school journal," *Journal of the American Society for Information Science*, vol. 40, no. 5, pp. 334–341, 1989.

[17] P. Mongeon and V. Larivière, "Costly collaborations: The impact of scientific fraud on co-authors' careers," *Journal of the Association for Information Science and Technology*, vol. 67, no. 3, pp. 535–542, 2016.

[18] J. Travers and S. Milgram, "The small world problem," *Phychology Today*, vol. 1, pp. 61–67, 1967.

[19] E. Rahm and A. Thor, "Citation analysis of database publications," *ACM Sigmod Record*, vol. 34, no. 4, pp. 48–53, 2005.

[20] E. A. Leicht, G. Clarkson, K. Shedden, and M. E. Newman, "Large-scale structure of time evolving citation networks," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 59, no. 1, pp. 75–83, 2007.

[21] X. Shi, L. A. Adamic, B. L. Tseng, and G. S. Clarkson, "The impact of boundary spanning scholarly publications and patents," *PloS one*, vol. 4, no. 8, p. e6547, 2009.

[22] I. Spiegel-Rosing, "Science studies: Bibliometric and content analysis," *Social Studies of Science*, vol. 7, no. 1, pp. 97–113, 1977.

[23] H. G. Small, "Cited documents as concept symbols," *Social studies of science*, vol. 8, no. 3, pp. 327–340, 1978.

[24] A. Athar and S. Teufel, "Context-enhanced citation sentiment detection," in *Proceedings of the 2012 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, pp. 597–601, Association for Computational Linguistics, 2012.

[25] A. Athar and S. Teufel, "Detection of implicit citations for sentiment detection," in *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pp. 18–26, Association for Computational Linguistics, 2012.

[26] S. Teufel, A. Siddharthan, and D. Tidhar, "Automatic classification of citation function," in *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 103–110, Association for Computational Linguistics, 2006.

[27] S. Ghosh, D. Das, and T. Chakraborty, "Determining sentiment in citation text and analyzing its impact on the proposed ranking index," in *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 292–306, Springer, 2016.

[28] S. Bonzi, "Characteristics of a literature as predictors of relatedness between cited and citing works," *Journal of the American Society for Information Science*, vol. 33, no. 4, pp. 208–216, 1982.

[29] A. Abu-Jbara, J. Ezra, and D. R. Radev, "Purpose and polarity of citation: Towards nlp-based bibliometrics.," in *HLT-NAACL*, pp. 596–606, 2013.

[30] A. Athar, "Sentiment analysis of citations using sentence structure-based features," in *Proceedings of the ACL 2011 student session*, pp. 81–87, Association for Computational Linguistics, 2011.

[31] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[32] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in *Proceedings of the 14th international conference on World Wide Web*, pp. 342–351, ACM, 2005.

[33] A. N. Langville and C. D. Meyer, *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2011.

[34] W. Webber, A. Moffat, and J. Zobel, "A similarity measure for indefinite rankings," *ACM Transactions on Information Systems (TOIS)*, vol. 28, no. 4, p. 20, 2010.