



ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com



Assessing Citation Relevance And Polarity Through A Combination of Semantic And Syntactic Information

¹Sangeetha Chamundeswari and ²G.S. Mahalakshmi

¹Department of Information Technology., Jawahar Engineering College, Affiliated to Anna University, Chennai 600093, Tamilnadu, India.

²Department of Computer Science and Engineering, Anna University , Chennai 600025, Tamilnadu, India.

ARTICLE INFO

Article history:

Received 10 November 2015

Accepted 30 December 2015

Available online 18 January 2016

Keywords:

Bibliometrics, Citation count

Citation classification Sentiment

analysis

ABSTRACT

Measurement of quality of research articles has received much attention in recent years. Number of citations that papers receive has become prominent in measuring research quality. Authors cite irrelevant articles to increase the article's popularity. Incorrect citation practices can increase the citation count and give unfair credit to research papers. In this paper we propose a novel automated technique which first classifies whether the cited article is sensible or not using a similarity method that integrates the semantic relations between words, and their syntactic composition. Then relevant citations are classified as sentiment positive, sentiment negative or neutral by identifying the cue phrases around the cited area. The proposed method is evaluated on articles collected from reputed open access journals on informatics. Results show that integrating semantic and syntactic approach produces better results than the existing methods. The proposed approach can be generalized to analyse research articles from various journals

© 2015 AENSI Publisher All rights reserved.

To Cite This Article: Sangeetha Chamundeswari., Mahalakshmi G.S, Assessing Citation Relevance And Polarity Through A Combination of Semantic And Syntactic Information. *Aust. J. Basic & Appl. Sci.*, 9(36): 340-348, 2015

INTRODUCTION

Bibliometrics is the science of analyzing statistically written publications, such as books or articles. Bibliometric methods are used mainly in the field of library and information science, which includes informatics and scientometrics. Bibliometrics is also used to provide quantitative analysis of academic literature. Citation analysis and content analysis comes under bibliometric methodologies. Many research fields use bibliometric methods to explore the impact of their field, the impact of a set of researchers, or the impact of a particular paper. It has other applications, such as the development of thesauri, descriptive linguistics and evaluation of reader usage.

A citation is a reference to a published or unpublished source. Citations are an abbreviated alphanumeric expression inserted in the body of an intellectual work for the purpose of acknowledging the relevance of the works done by others regarding the topic being discussed at the spot where the citation appears.

Citation analysis deals with the examination of the documents cited in research works. Its main application was originally information retrieval and

analyzing the quality of research papers. Recent works show that they are used in bibliometrics for evaluating and mapping research works, measuring the production and dissemination of scientific knowledge and also for establishing the journal impact factor. Citation indices are the main source for citation analysis which is bibliographic databases that allow one to establish citation details such as which later documents cite which earlier documents, which articles have been cited most frequently and who has cited them.

Measuring quality of research is one of the trending topics in recent times. A number of bibliometric indicators are used to assess the quality of research outputs in the higher education sector worldwide. The most common among these are indicators are the one based on journal impact factors, journal ranks and citations. Details of citation such as the presence and number of citations are frequently used to assess the influence of a particular article, author, journal or field of research. But it is registered that the numbers of citations do not necessarily correlate with article quality by Todd, Peter A and *et al.*,(2010).

Studies have found that citation behavior is complex in itself. Why an article is cited or not cited

Corresponding Author: V. Sangeetha Chamundeswari, Department of Information Technology., Jawahar Engineering College, Affiliated to Anna University, Chennai 600093, Tamilnadu, India.
E-mail: sangeetharam26@gmail.com

is dependent on several factors. The advantage of using metrics for measuring quality of research papers is that they are objective and can be measured easily. However metrics alone cannot be completely relied on for measuring its relevance. Part of this challenge lies within using the citation index to capture research outputs.

A number of factors are involved to determine the citation of a paper. Citations depend on time, journal, field or authors. Time feature includes how recently the work has been done. The influential factors in journal are its prestige, accessibility, visibility and internationality. Field dependent features include specificity of topic and field of study whereas the author/reader dependent features include language of the article, social networks of the author, number of co-authors and prestige of the author.

Articles which are contentious or debatable also attract a lot of citations even though they may contribute little to science or might be cited due to methodological problems and are known as negative citations. These improper citations give unfair credit to research papers. Thus there is a need for a new method to eliminate irrelevant citations and a new technique to assess the quality of research papers.

Literature Survey:

The research accomplishment of individuals, groups and institutions are increasingly being quantified using bibliometric-based performance indicators referred by Todd, Peter A and Ladle, Richard J, Citations., (2008) and Adler, Nancy J and Harzing, Anne-Wil., (2009). Such metrics are popular because they are relatively objective, transparent and they are also quick and easy to calculate. They have their downside as undeserving paper gets unnecessary boost in their citation counts which have no reflection on research to be credited.

Todd, Peter A *et al.*, (2007) was the first to measure citation misconduct in the branch of ecology. They determined that approximately one-quarter of citations were ambiguous, empty (citations to secondary sources), or did not support the assertion at all.

Dupps, William J (2008) comments that citation is subjective to a human ideas and papers are not referred simply on academic appropriateness or merit by Bornmann *et al.*, (2008). Numerous factors are known to affect the probability of a paper being cited, including the language used, the number of authors by Della Sala *et al.*, (2008) plus their affiliations and status by Leimu, *et al.*, (2009), the paper's length and the significance of the results. There are also multiple issues in determining the precise number of cites a paper has accrued by Stergiou, Konstantinos I and Tsikliras, Athanassios.,(2006).

Some proactive measures are carried out to address citation malpractice, which includes authors signing a declaration stating their citations have been

verified by Goldberg R *et al.*,(1993), or publishing errors that are spotted by readers by De Lacey *et al.*,(1985). Random checking of references for a new submission is also considered. If errors are found, the manuscript can be returned and the author asked to provide the relevant parts of all work cited in their paper. Any such procedure that might result in time-to-publish penalties or outright rejection should provide sufficient incentive for authors to check their work for citation accuracy.

Sentiment analysis of citations in scientific papers and articles is a new and interesting problem due to the many linguistic differences between scientific texts and other genres. Teufel, Simone *et al.*,(2006) worked on a 2829 sentence citation corpus using a 12-class classification scheme. The corpus has been annotated for the task of determining the author's reason for citing a given paper and is thus built on top of sentiment of citation. The 12 classes were grouped into 3 categories in an attempt to perform a rough approximation of sentiment analysis over the classification.

Todd, Peter A and Ladle, Richard J.,(2008) proposed a novel automated technique, which classifies whether an earlier work is cited as sentiment positive or sentiment negative. This approach first extracted the portion of the cited text from citing paper. Using a sentiment lexicon they classified the citation as positive or negative by picking a window of at most five sentences around the cited place. The result showed that 80 per cent of the citations are positive and 20 per cent of the citations are negative.

Citing sources that are not relevant when cited only mislead the readers. It also rips the credit going to the actual contributors of work. To quantify citation fidelity in marine biology, Todd, Peter A *et al.*, (2010) retrieved 198 papers from 2 recent issues of 33 marine biology journals and evaluated its appropriateness. It was discovered that the assertion was clearly supported by the citation in only 75.8% of cases, the support was ambiguous in 10.6% of cases and the citation offered no support to the original statement in 6.0% of cases. The remaining 7.6% of cases were classified as empty (citations to secondary sources). It was found that 1 in 4 citations in marine biology are inappropriate.

Abdi, Asad and Idris *et al* (2015) proposed a method that integrates the semantic relations between words and their syntactic composition to find similarity between a sentence and a document. Database and sentence structure act as sources for semantic and syntactic processing. Similarity measure between two sentences is obtained using a linear equation that combines the semantic and word order similarity. As a result, the proposed method is able to obtain high accuracy and improve the performance compared with the current techniques.

The particle swarm optimization (PSO) algorithm is a search algorithm that concentrates on

population based stochastic information. the algorithm was inspired by the social behaviour of bird flocks or schools of fish. since its introduction in 1995, pso has drawn much attention and has been applied to the solution of optimization problems by Poli, Riccardo and Kennedy *et al.*,(2007).

Therefore finding the relevancy and sentiment of cites is necessary for evaluating research articles. A combination of semantic and syntactic similarity measure can be used to find the relevancy of cite. Sentiment of a citation could be identified with the cue phrases that are annotated to the cite.

Methodology:

The proposed first classifies whether the cited article is sensible or not using a similarity method that integrates the semantic relations between words, and their syntactic composition. Then relevant citations are classified as sentiment positive, sentiment negative or neutral by identifying the cue phrases around the cited area. The overall architecture of the proposed work is shown in Figure 1.

Data collection:

Articles for analysis are collected from reputed open access journals on info metrics. It consists of 9 volumes and 4 Issues per volume out of which 2700 articles are available for open access. These articles are downloaded using a crawler. The reference articles for the article analyzed are manually downloaded in pdf format

Data conversion and extraction:

Root paper and all its citing papers are indexed. First step is to convert all pdf files in dataset into text format. Pdf-to-text conversion tends to modify formatting style, making it difficult to parse the text. Individual sentences are parsed and noisy data is removed. Based on the format style of the citing paper a particular regular expression is executed to extract the reference. With the help of this reference we form another regular expression to extract cite of maximum 150 words from the citing paper. The extracted cite is used for further processing. A sample of the processed corpus is shown in Table 1.

Table 1: Pre-processed documents

Pre-processed corpus	Pre-processed source
C1. General form citation distributions heavily skewed small percentage highly cited papers large proportion low cited	A1. Paper studies evidence thomson scientific citation process articles published period web science was categories sub fields A2. Reference citation distributions characteristics sub fields A3. Analyzed characteristic scores scales css technique replication scale invariant shape distributions three broad categories articles appears strikingly A4. Reference distributions mildly skewed citation distributions year citation window highly skewed points median articles upper tail account citations A5. Aggregation sub fields disciplines fields aggregation schemes preserve feature citation distributions A6. Subsets articles lower upper tails citation distributions universality partially breaks A7. Hand sub fields existence power law rejected A8. Contrary generally believed sub field level scaling parameter time power laws small average represent articles account citations

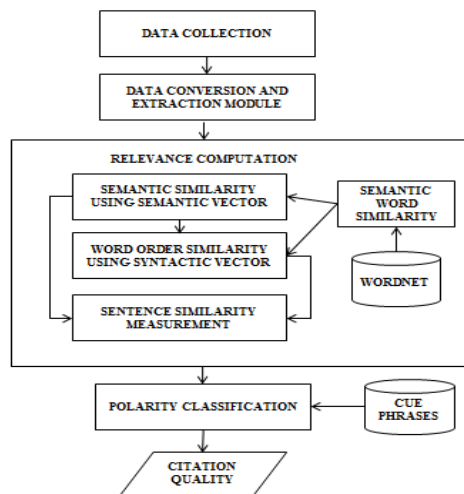


Fig. 1: Overall architecture

Relevance computation:

The cite text and the corresponding article are decomposed into a set of sentences. Then, the similarity measures between each sentence from cite

and whole sentences from the source article text are determined using the composition of word order similarity and semantic similarity as shown in Figure 2.

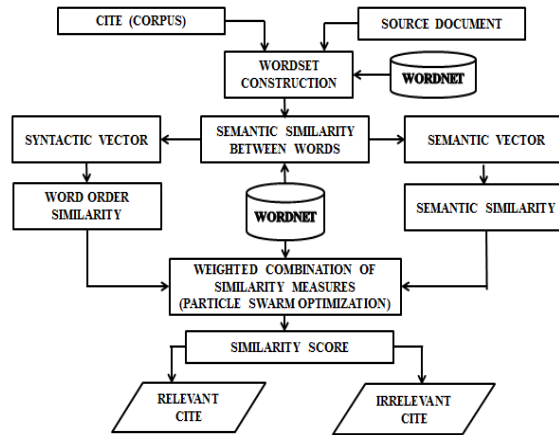


Fig. 2: Relevance computation

Word set construction:

Given two sentences from source document and corpus document, a word set $WS = \{WS_1, WS_2, \dots, WS_N\}$ is created using distinct words from the two sentences, where N is the number of distinct words in word set. For each word W from sentence1, the root

word of the word W is obtained through WordNet. If the root word does not appear in the WS, it is added to word set WS. The same process is repeated for Sentence 2. A sample word set for a pair of sentence is shown in Table 2.

Table 2: Word set

For pair c1a4	Word set
c1 general form citation distributions heavily skewed small percentage highly cited papers large proportion low cited a4 reference distributions mildly skewed citation distributions year citation window highly skewed points median articles upper tail account citations	[general form citation distributions heavily skewed small percentage highly cited papers large proportion low reference mildly year window points median articles upper tail account citations distribution skew cite paper point article]

Semantic similarity between words:

Semantic word similarity is used to create word order vector and semantic vector. Given two words W1 and W2, the semantic similarity between words is determined with the following steps:

1. Get root of each word using Word Net
2. Get synonym of each word using Word Net
3. Calculate the number of synonyms of each word
4. Calculate the Least Common Subsume (LCS) of two words and their length
5. Determine the similarity score between words using Eq. 1 and 2

$$IC(w) = 1 - \frac{\log(\text{synset}(w)+1)}{\log(\text{max}_w)} \tag{1}$$

$$\text{sim}(w_1, w_2) = \begin{cases} \frac{2 \cdot IC(LCS(w_1, w_2))}{IC(w_1) + IC(w_2)}, & \text{if}(w_1 \neq w_2) \\ 1, & \text{if}(w_1 = w_2) \end{cases} \tag{2}$$

where LCS stands for the least common subsume, max_w is the number of words in Word Net, Synset(w) is the number of synonyms of word w, and IC(w) is the information content of word w based on the lexical database Word Net. A sample for word order similarity is shown in Figure 3.

$$\begin{aligned}
 \text{LCS (ARTICLE,CITATION)} &= \\
 \text{communication} &= \\
 \text{SIM(ARTICLE,CITATION)} &= \\
 \frac{2 \times \text{IC (COMMUNICATION)}}{\text{IC (ARTICLE)} + \text{IC (CITATION)}} &= \frac{2 \times 4}{9 + 9} \\
 &= 0.4210526
 \end{aligned}$$

Fig. 3: Similarity between words

Semantic similarity between sentences:

The semantic-vector approach is used to measure the semantic similarity between sentences. The semantic vector is constructed from the word set and the corresponding sentence.

The dimension of semantic vector equals the number of words in the word set. Each cell of the semantic vector corresponds to a word in the word set, and its weight is calculated using the semantic similarity between words from the word set and corresponding sentence. If the word W from the word set appears in the sentence S1, the weight of word W in the semantic vector is set to 1. If the word W does not appear in the sentence S1, then the similarity between the word W and all of the words in the sentence S1 is calculated as shown in Eq. 1 and 2. The weight of word W in the semantic vector is set to the highest similarity value. If there is no similarity value between word W and all of the words in the sentence S1, the weight of word W in the semantic vector is set to 0. A semantic-vector is constructed for each of the two sentences. A sample semantic vector is shown in Table 3. The semantic

similarity is computed using the two semantic vectors. Eq. 3 is used to calculate the semantic similarity between sentences.

$$\text{Sim}_{\text{semantic}}(S_1, S_2) = \frac{\sum_{j=1}^m (w_{1j} \times w_{2j})}{\sqrt{\sum_{j=1}^m w_{1j}^2} \times \sqrt{\sum_{j=1}^m w_{2j}^2}} \quad (3)$$

where $S_1 = (w_{11}, w_{12}, \dots, w_{1m})$ and $S_2 = (w_{21}, w_{22}, \dots, w_{2m})$ are semantic vectors of S_1 and S_2 respectively.

Word order similarity between sentences:

The syntactic-vector approach is used to measure word order similarity between sentences. The syntactic vector is constructed from the word set and the corresponding sentence.

The dimension of syntactic vector equals the number of words in the word set. Each cell of the syntactic vector corresponds to a word in the word set, and its weight is calculated using the semantic similarity between words from the word set and corresponding sentence.

Table 3: Semantic vector

Word set	Semantic vector
[general form citation distributions heavily skewed small percentage highly cited papers large proportion low reference mildly year window points median articles upper tail account citations distribution skew cite paper point article]	C1. General form citation distributions heavily skewed small percentage highly cited papers large proportion low cited 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 0.47999998927116394 0.0 0.6190476417541504 0.5789473652839661 0.4000000059604645 0.5531914830207825 0.4285714328289032 0.523809552192688 0.5 0.5454545617103577 0.5652173757553101 0.5777778029441833 0.0 0.4285714328289032 0.5925925970077515 0.4000000059604645 0.5

If the word W from the word set appears in the sentence S1, the weight of word W in the syntactic vector is set to the index position of the word in sentence. If the word W does not appear in the sentence S1, then the similarity between the word W and all of the words in the sentence S1 is calculated as shown in Eq. 1 and 2. The weight of word W in the syntactic vector is set to the index position of the word with highest similarity value. If there is no similarity value between word W and all of the words in the sentence S1, the weight of word W in the syntactic vector is set to 0. A syntactic-vector is

constructed for each of the two sentences. A syntactic-vector is constructed for each of the two sentences. A sample semantic vector is shown in Table 4. The word order similarity is computed using the two syntactic vectors. Eq. 4 is used to calculate the semantic similarity between sentences

$$\text{Sim}_{\text{word order}}(S_1, S_2) = 1 - \frac{\|O_1 - O_2\|}{\|O_1 + O_2\|} \quad (4)$$

where $O_1 = (d_{11}, d_{12}, \dots, d_{1m})$ and $O_2 = (d_{21}, d_{22}, \dots, d_{2m})$ are syntactic vectors of S_1 and S_2 respectively

Table 4: Syntactic vector

Word set	Semantic vector
[General form citation distributions heavily skewed small percentage highly cited papers large proportion low reference mildly year window points median articles upper tail account citations distribution skew cite paper point article]	C1. General form citation distributions heavily skewed small percentage highly cited papers large proportion low cited 1 2 3 4 5 6 7 8 9 10 11 12 13 14 11 1 0 0 6 4 11 9 7 0 3 4 6 10 11 6 11

Weighted combination of the similarity measures:

The similarity measure between two sentences is calculated using a linear equation that combines the semantic and word order similarity. The similarity measure is computed as shown in Eq. 5

$$sim_{sentences}(S_1, S_2) = \lambda \cdot sim_{semantic}(S_1, S_2) + (1 - \lambda) \cdot sim_{wordorder}(S_1, S_2) \quad (5)$$

where λ is the weighting parameter, that specifies the relative contributions to the overall similarity measure from the semantic and syntactic similarity measures. λ ranges from 0 to 1. If $\lambda = 0.5$ the semantic and syntactic similarity measures are assumed to be equally important. Eq. 5 is optimized using Particle Swarm Optimization (PSO). Similarity values input to the PSO is as shown in Table 5.

Table 5: Similarity values

Word order similarity	Semantic similarity
c1 a1 0.206641565177016	c1 a1 0.4164405149780775
c1 a2 0.5087547913309824	c1 a2 0.7391669057915946
c1 a3 0.328613779096192	c1 a3 0.4947681187271748
c1 a4 0.7387286431453528	c1 a4 0.8836725557637015
c1 a5 0.565017219573208	c1 a5 0.7269355252783872
c1 a6 0.5313005820517517	c1 a6 0.7264994425366171
c1 a7 0.1182012347429651	c1 a7 0.3206218143060582
c1 a8 0.5454545454545454	c1 a8 0.6986612815647122
c1 a9 0.5162205531531032	c1 a9 0.6858575460554278

Particle Swarm Optimization:

PSO is a computational method for optimizing (maximizing or minimizing) a problem which tries to improve a candidate solution in each iteration with regard to a given measure of quality.

PSO uses a swarm of N_{sw} particles and each particle i of the swarm has a position in the continuous n dimensional search space.

$$p_i(t) = [p_{i1}(t), p_{i2}(t), \dots, p_{in}(t)] \quad (6)$$

At each iteration, particle's best position and velocity is calculated and recorded as

$$p_i^{best} = [p_{i1}^{best}, \dots, p_{in}^{best}] \quad (7)$$

$$v_{ij}(t) = [v_{i1}(t), \dots, v_{in}(t)] \quad (8)$$

The personal best position of each particle at iteration $(t+1)$ is calculated using Eq.5.

At each iteration, global best position of the swarm is computed as in Eq.9.

$$g^{best}(t) = \max \{ f(p_1(t)), \dots, f(p_{N_{sw}}(t)) \} \quad (9)$$

The initial position and velocity is generated using Eq.10 and Eq.11.

$$p_{ij}(0) = p_{min} + (p_{max} - p_{min}) \cdot r_1 \quad (10)$$

$$v_{ij}(0) = v_{min} + (v_{max} - v_{min}) \cdot r_2 \quad (11)$$

where v_{min} and v_{max} are the minimum and maximum allowed velocity and where p_{min} and p_{max} are the minimum and maximum allowed position of the particle. r_1 and r_2 are two independent random numbers uniformly distributed within the interval $[0,1]$.

At each iteration, particle's position and velocity is updated using Eq.12 and Eq.13.

$$v_{ij}(t+1) = w \cdot v_{ij}(t) + c_1 \cdot r_1 \cdot (p_{ij}^{best}(t) - p_{ij}(t)) + c_2 \cdot r_2 \cdot (g_j^{best}(t) - p_{ij}(t)) \quad (12)$$

$$p_{ij}(t+1) = p_{ij}(t) + v_{ij}(t+1) \quad (13)$$

where w is the inertia weight that controls the influence of previous velocity on the current velocity and is calculated using Eq.14.

$$w(t) = \frac{(t_{max} - t) \cdot (w_{max} - w_{min})}{t_{max}} \quad (14)$$

where t is the current iteration, t_{max} is the maximum number of iterations, w_{max} and w_{min} are the starting and ending inertia weight values to control inertia.

An intermediate parameter than controls c_1 and c_2 is calculated as in Eq.15.

$$\phi = ((w+1) \cdot (w+1)) / (2 \cdot w) \quad (15)$$

A sample particle position is shown in Table 6. Each particle has 'n' dimensions where each dimension stands for a pair of sentence formed between the source document and corpus. The dimensions will be initialized randomly in the first iteration as shown in Eqn.10. At each iteration the particle position will be updated using a velocity factor as shown in Eqn.12 and Eqn.13. Higher the value of a dimension in a particle denotes that the similarity between the corresponding pair of sentences is high.

Table 6: Particle position

Position of 5 particle					
n0	0.7234619346475598	0.7455251247163597	0.570080884902278	0.9761345009479563	0.9093087645943803
	0.89489206252175162	0.2933047762244807	0.71238482139203824	0.775701495205073	
n1	0.70461622775399748	0.9645898366476167	0.5702926497593203	0.8266843367661946	0.43810230625551694
	0.2776473263615432	0.3467114322488909	0.8833208349793101	0.15589303315321645	
n2	0.036956148206992	0.844736817251778	0.8892532581519116	0.80363613004500565	0.44587692393580145
	0.6910187081391244	0.41648476857146255	0.9374387312502569	0.02822434255707784	
n3	0.22159849609644777	0.34294608964804896	0.3160655268072897	0.50693157470008345	0.4603323591284786
	0.9939397356381856	0.5558930629392004	0.46539345775012364	0.33538204388772763	
n4	0.10826021382944662	0.29166275963693283	0.1213929866405286	0.34547036399074113	0.4876557865574714
	0.1515126695649606	0.655267950486914	0.6612005874742439	0.9992060859209719	

Table 7: Fitness values

Fitness value
pbest0= 1.4034770671811325
pbest1= 0.9064468734123491
pbest2= 0.8958459795155738
pbest3= 0.4579094971172409
pbest4= 0.4441747877665999
gbest=1.4034770671811325

Table 7 shows the fitness value of the particles. High values of pbest denote that the particle has found the more accurate results. The best particles fitness in a particular iteration is stored in gbest. From Table 6 and Table 7 it can be inferred that particle 'n0' has found the more accurate results and the fourth pair of sentence (c1a4 according the example in Table 1) is more relevant.

Polarity classification:

Based on the cue phrases citations are classified into 12 categories [4] as shown in Table.8

To perform sentiment analysis the 12 categories are grouped into 3 categories as shown in Table. 9.

Table 8: Citation classification categories

Category	Description
Weak	Weakness of cited approach
CoCoGM	Contrast/Comparison in Goals or Methods(neutral)
CoCo-	Author's work is stated to be superior to cited work
CoCoR0	Contrast/Comparison in Results (neutral)
CoCoXY	Contrast between 2 cited methods
PBas	Author uses cited work as basis or starting point
PUse	Author uses tools/algorithms/data/definitions
PModi	Author adapts or modifies tools/algorithms/data
PMot	This citation is positive about approach used or problem addressed (used to motivate work in current paper)
PSim	Author's work and cited work are similar
PSup	Author's work and cited work are compatible/provide support for each other
Neut	Neutral description of cited work, or not enough textual evidence for above categories, or unlisted citation function

Table 9: New Categories

Old categories	New categories
Weak, CoCo-	Negative
PMot, PUse, PBas, PModi, PSim, PSup	Positive
CoCoGM, CoCoR0, CoCoXY, Neut	Neutral

Evaluation:**Evaluation techniques:**

In order to make inference about our proposed method we picked random citations and evaluated manually. The results are quantified using the following metrics.

Precision:

Precision determines the exactness of classification i.e., the percentage of tuples that the classifier labeled as positive is actually positive.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall:

Recall determines the completeness of classification i.e., the percentage of positive tuples the classifier labeled as positive

$$\text{Recall} = \frac{TP}{TP+FN}$$

F-measure:

– F-score determines the harmonic mean of precision and recall.

$$F = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Performance analysis:

The relevance computation phase showed that 7.3% of the references are not cited. Out of the remaining 92.7% citations 76% citations are relevant, 14% citations are ambiguous (partial support) and 10% are irrelevant.

The results of sentiment analysis of the cited corpus are shown in Table 10.

Table 10: Results of evaluation

Measures	Value
Precision	0.84
Recall	0.94
F-Score	0.89

The precision and recall values were improved considerably using this method. Using relevancy and sentiment of cite we can further check whether the earned citation of a research paper is worthy. By constructing a citation graph and pruning it by removing the irrelevant citation edges, the implicit and explicit citation count earned by a research paper can be calculated.

Conclusion:

We have proposed a novel automated technique which first classifies whether the cited article is relevant or not using a similarity method that integrates the semantic relations between words, and their syntactic composition. Then relevant citations are classified as sentiment positive, sentiment negative or neutral by identifying the cue phrases around the cited area. The proposed method is evaluated on articles collected from reputed open access journals on informatics. The proposed approach is evaluated using precision, recall and f-measure. The proposed method could be used by research scholars and expert panels for evaluating the journal articles.

REFERENCES

Abdi., Asad and Idris., Norisma and Alguliev, M Rasim and Aliguliyev, M. Ramiz, 2015. Automatic summarization assessment through a combination of semantic and syntactic information for intelligent educational systems. *Journal of Information Processing & Management*, 51(4): 340-358.

Adler, Nancy J and Harzing, Anne-Wil, 2009. When knowledge wins: Transcending the sense and

nonsense of academic rankings, *Journal of Academy of Management Learning & Education*, 8(1): 72-95.

Bornmann., Lutz and Daniel., Hans-Dieter, 2008. What do citation counts measure? A review of studies on citing behaviour, *Journal of Documentation*, 64(1): 45-80.

Butt, Bilal Hayat and Rafi, Muhammad and Jamal, Arsal and Rehman, Raja Sami Ur and Alam, Syed Muhammad Zubair and Alam, Muhammad Bilal, 2008. Classification of research citations, *Journal of arXiv*, 1506(8966).

De Lacey, Gerald and Record, Carol and Wade, Jenny, 1985. How accurate are quotations and references in medical journals?, *British Medical Journal*, 291(6499):884-886.

Della Sala, Sergio and Brooks, Joanna, 2008. Multi-authors' self-citation: a further impact factor bias? *Journal of Cortex*, 44(9): 1139-1145.

Dupps, William J. 2008. Impact of citation practices: beyond journal impact factors, *Journal of Cataract Refractive Surgery*, 34(9): 1419-1421.

Goldberg, R., E. Newton, J. Cameron, R. Jacobson, L. Chan, W.R. Bukata, A. Rakab, 1993. Reference accuracy in the emergency medicine literature. *Journal of Annals of emergency medicine*, 22(9): 1450-1454.

Leimu, Roosa and Koricheva, Julia, 2005. What determines the citation frequency of ecological papers?, *Journal of Trends in Ecology & Evolution*, 20(1): 28-32.

Poli, Riccardo and Kennedy, James and Blackwell, Tim, 2007. Particle Swarm Optimization: An Overview, *Journal of Swarm Intelligence*, 1(1): 33-57.

Stergiou, Konstantinos I and Tsikliras, Athanassios, 2006. Underrepresentation of regional

ecological research output by bibliometric indices, *Journal of ethics and environment*, pp: 15-17.

Teufel, Simone, Advaith Siddharthan, and Dan Tidhar, 2006. Automatic classification of citation function, *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp: 103-110.

Todd, Peter A and Guest, James R and Lu, Junxiu and Chou, Loke Ming, 2010. One in four citations in marine biology papers is inappropriate,

Journal of Marine Ecology Progress Series, 408: 299-303.

Todd, Peter A and Ladle, J. Richard, 2008. Citations: poor practices by authors reduce their value, *Journal of Nature*, 451(7176): 244-254.

Todd, Peter A and Yeo, DC and Li, Daiqin and Ladle, J. Richard, 2007. Citing practices in ecology: can we believe our own words?, *Journal of OIKOS-COPENHAGEN*, 116(9): 1599.