CITATION SEMANTIC BASED APPROACHES TO IDENTIFY ARTICLE QUALITY

S. Sendhilkumar[1], E. Elakkiya[2] and G.S. Mahalakshmi[3]

[1]Department of Information Science and Technology
Anna University, Chennai, Tamil Nadu
ssk_pdy@yahoo.co.in
[2]Department of Computer Science and Engineering
Anna University, Chennai, Tamil Nadu
elakkiyae@gmail.com
[3]Department of Computer Science and Engineering
Anna University, Chennai, Tamil Nadu
mahalakshmi@cs.annauniv.edu

ABSTRACT

Analyzing the structure of research articles and the relationship between the articles has always been important in bibliometrics research area. One of the methods for analyzing articles is by its citation. Normally researchers rank the article based on the citation count. But this quantity based evaluation is not sufficient due to various citation types like Random citation, Guest citation. Our paper aims in providing a new method to rate a research paper based on its citation quality. In order to find the citation quality, three different semantic related processes are used: Citation classification, Citation Sentiment Analysis, Content Relevancy. This work analyzes some challenges found in citing a research paper. We have proposed methods to figure out the valid citations for research paper and thereby found quality of the research paper.

KEYWORDS

Citation Analysis, Citation classification, Citation Extraction, Citation Networks, Citation sentiment analysis.

1. INTRODUCTION

In the research field, there are different perspectives to view the quality of journals. Some of them are trusted content of journal, research and data correctness, proper level of research, author trustworthy and article quality. Journal quality is measured by impact factors. “The impact factor is calculated by dividing the number of current citations to articles published in the two previous years by the total number of articles published in the two previous years." (Garfield, 1999). This refers to the amount of contribution of a particular journal to a research area. Citations are the process to acknowledge the source of the author used in the published work. Number of citations

David C. Wyld (Eds) : ICCSEA, SPPR, CSIA, WimoA, SCAI - 2013
pp. 411–420, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3543

is counted and this numeric value is used as a metric to calculate the impact factor. The measure thus obtained is further used to assess the quality of:

- An individual article quality - measured by how often it was cited
- An author quality - measured by total citation or average citation count per article.
- A journal quality - measured by average citation count for the article in the journal

Journal quality is measured by article quality and author quality (Deepika et al., 2011). All articles published in a journal are not of same quality. The originator of the impact factor (Garfield, 1999) also states that it is incorrect to judge an article by the impact factor of the journal because citations are not uniformly distributed between articles. There is a vital need to find a different method which will identify article quality and thereby select a valid reference.

This work proposes a new method to identify article quality by semantic based techniques which analyze the citation sentences by: 1) focusing on the aspects in which a cited article cites the reference article, 2) analyzing the sentiment of citation like positive or negative and 3) identifying content relevancy which will yield the correct quality score. Our experiments confirm the effectiveness of the method and show that it outperforms other state of the art quantitative techniques.

2. RELATED WORK

Various types of evaluation have been made based on the citation. Initially, frequency of citations received from Science Citation Index (SCI) database about the article was used to find the journal quality (Garfield, 1999). SCI ranks the journal based on the number of citation an article receives. If an article is cited less frequently it is given lesser reputation even if the quality of content in the article is good.

Later with the use of Graph-theoretic approaches for ranking network entities, researchers have moved to introduce link analysis approaches. PageRank (Page et al., 1998) was used for citation counting. Here the citations are considered to be in a link structure and the citations are ranked based on the number of forward (outgoing) links and backward (incoming) links an article has and the importance of nodes from and to which the link flows. But PageRank is mainly used for web pages than Research article. Hyper text Induced Topic Search (HITS) (Kleinberg, 1999) was later used for ranking. It is very similar to PageRank, except that it creates two popularity score instead of one and it considers both in links and out links to create popularity scores for each page.

Comprehensive Citation Index (CCI) (Bi etal., 2011) considers both direct and indirect influence of research article on its citing papers through citation links on even those papers that do not directly cite it. Such indirect influence decreases for each citation link. These methods are based on quantitative measures. Heterogeneous PageRank algorithm (Lagville et al, 2006) was used later. This algorithm is based on the assumption that - there would be a different propagation probability for a node to follow different kinds of out-going links (links to different types of nodes).

Citation Classification is concerned with identifying the nature of connection between the cited and citing articles. The earliest citation scheme lists the reasons why authors cite other works

(Garfield, 1965). The first classification of citation divides citations in running text into four dimensions rather than one classification function (Moravcsik and Murugesan, 1975) namely: conceptual or operational use , evolutionary or juxtaposition, organic or perfunctory and confirmative or negational . Another scheme classifies citations into Seven Argumentative Zones say, *Background*, *Other*, *Own*, *Aim*, *Textual*, *Contrast*, and *Basis*, according to their role in the author's argument (Teufel et al., 1999). A completely diverse yet simple classification was proposed which composed of only three categories namely Type B, C and O (Nanba and Okumura 1999, 2000). Another classification of citation consist 12 category framework based on the empirical work in citation content analysis (Simone Teufel, 2006). The classification are Weakness of cited approach, Contrast Comparison in Goal & Results, base, uses, modifies, motivate, similar, support and neutral. Yet another classification scheme (Pham et al, 2003) classifies citations into 4 categories, such as Basis, Support, Limitation and Comparison. Using Ripple Down rules citation context were categorized into these category.

In most of the papers sentiment of citation were discussed within the citation classification. But the reason why we present here separately is the importance of the topic. Automatic identification of sentiment polarity in citations represent each citation as a feature set in SVM framework and the author argues that it produces good results for sentiment classification (Athar et al., 2011). Sentiment analysis was used to rate citations as positive, neutral or negative along with the help of a Lexical Analyzer called SentiWordNet (Diana et al., 2011).

3. METHODOLOGY

The framework for extracting quality score is given in the following steps and it is depicted in figure 3.1

1. Retrieve the citation article from the Google scholar automatically. This enables us to keep track of real time updates to citation articles.

2. Extract the citation context which may have one or more citation instances and cites.

3. Identify Sentiment of the article which enables us to identify whether the author says positive or negative things about the paper.

4. Classify the citation instances into different category for identifying author motivation about the cited article.

5. Identify the citation relevancy by cosine similarity and detect outliers. Citation outlier is the content of citation article that is relevant to the article but not greatly relevant to cited article.

6. Aggregate the citation score that retrieved from the above steps into one.

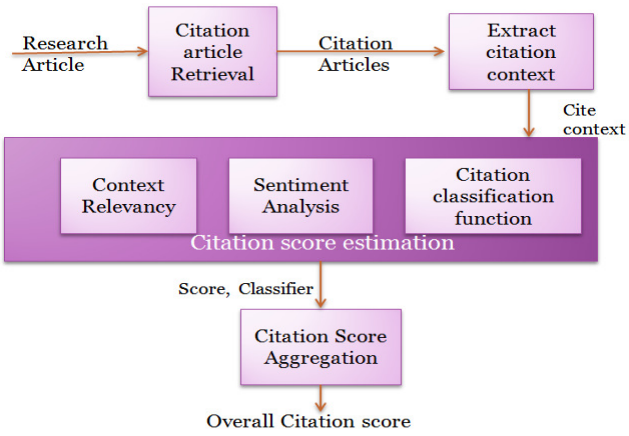

Fig 3.1: General Framework

Citation Retrieval

Initially the titles of seed articles are retrieved from Google Scholar. Later, search query is formed with every article title and submitted to search engine. This returns the query results with the link to citations to that article under 'cited by' category. Accessing the 'cited by' link enables us to retrieve the citation articles to that seed article. This web based retrieval enables us to monitor the real time update of citations to that article. However, restricted access of research articles results in reduced web based retrieval of actual citations.

Extracting Citation Context

Citation context refers to the sentences that speak about the cited article. Such sentences can be identified by the placeholders at the end of the sentence or anywhere within where the context of cited work is used. However extraction of citation context is very difficult due to various styles of citation references. Trained citation CRF file (Zhang 2009) is a probabilistic model with some learning feature. Using that CRF file, the citation context is segmented from the whole article and the full citation is parsed to recognize fields, including author name, title, and source.

Sentiment Analysis

Sentiment analysis of citations in research articles is a new and interesting problem as there are many linguistic differences between scientific texts and other genres. We used SentiWordNet (Baccianellaet al., 2010) in our work for this purpose to identify the sentiment of citation as positive, negative or neutral (Diana et al., 2011). SentiWordNet has 117,374 annotated synsets from WordNet 2.0 with sentiments scores. In SentiWordNet synsets are assigned with some numerical score with the notations Pos(s), Neg(s), Obj(s), which are the positivity, negativity and objectivity of the each synset respectively. The overall numerical score of the notation is equal to 1 distributed in the range from 0.0 to 1.0. The procedure to obtain the sentiment for the citation is as follows. The citation context is segmented into sentences. The sentences are then brought into being using part-of-speech tag as an annotation on each word or symbol. The sentiment score for each adjective is found from SentiWordNet Lexical Analyzer. All adjective scores are aggregated to obtain overall sentiment score. Adjectives are considered because mostly adjectives represent the sentiment in a sentence.

Classification

Citation Classification is concerned with identifying the nature of connection between the cited and citing articles. We use the classification scheme (Simone Teufel, 2006) that categories as compare contrast, basis, support, use, modifies, weak and simple. Sentences with existing citations are used as training data after removing the citation marker. For each paper from the dataset Training set is got from the examples annotated with class values. For each citation context the appropriate features were extracted and the classifier was constructed using Naïve Bayes algorithm. This classification has non numerical label.

Content Relevancy

The previous techniques that we discussed in this paper are based on the citation context retrieved from each citation article but this section focus on the full text of cited and citation article. Identifying the relevance of cite to a particular context is done by cosine similarity and outlier determination (Mahalakshmi et al., 2012).. Outlier determination is done so as to identify articles that are not greatly relevant to cited article. The citation outlier is found by Latent Dirichlet allocation (LDA) that is widely used for identifying the topics in a set of documents.The probability distribution is found based on Gibbs sampling and distribution of content over various topics is identified. The cited article and citation article are topic modeled using LDA and the distributions of two articles are identified. Then the similarity between the two topic distributions is computed. If they are at least 50% similar, then the citations are found to be apt, else the base paper is considered to be an outlier for that citation.

Aggregating the Citation Score

Each citation article may be referenced two or more times in the same paper. Each reference point is called the citation instances. Classification is non numerical values that can be used for the purpose of aggregating cite instance values because aggregation of quality score is based on the aspects of the citation. Citation quality score is calculated by sentiment score, similarity score, outlier score. The classification categories are ordered based on the importance as: Compare Contrast, Basis, Support, Use, Modifies, Weak and Simple. In the aggregation process the cite instance belonging to the highest ranking category is selected and its scores are aggregated.

$$Citation\ Quality\ Score = sentiment_i + Similarity_i + LDA_i \qquad eq.(1)$$

Where i =highest importance classification category.

In case of the Citation instance being an outlier then aggregation process omits the LDA similarity score.

$$Citation\ Quality\ Score = sentiment_i + Similarity_i \qquad eq.(2)$$

Where i =highest importance classification category.

4. RESULTS AND DISCUSSION

A set of documents are downloaded from the internet in the citation retrieval part and tested on the system. The Fig 4.1 depicts results obtained from the citation download for a given article. Main observations made include the total no of citation downloaded from the real citation count, the total no of article correctly downloaded out of the total citation downloaded and identify the Precision, Recall and F1 measure.

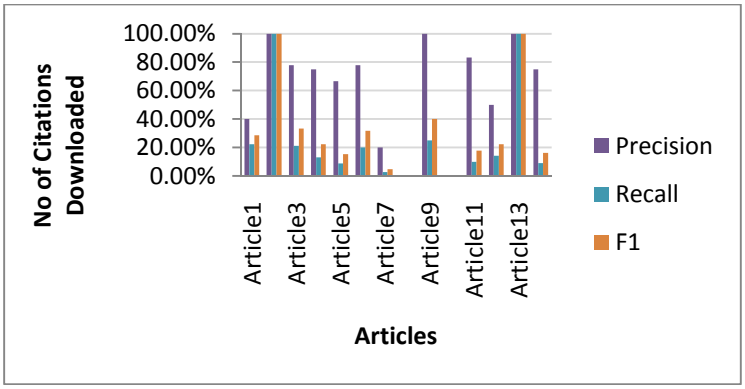

Fig 4.1 Precision, Recall, F1measure

In citation context extraction using supervised learning method we implemented training and test data. Precision, recall and F1 score are found based on the total number of annotated fields, total number of correctly identified field and total number of retrieved fields as shown in Table 4.1.

We evaluated the correctness of the result with manual examination. From that we identified some values are wrongly predicted like source. This is because of the ambiguity that occurred between the source, the title and author name.

Fields	Precision	Recall	F1
Title	100.00%	88.89%	94.12%
Source	62.50%	55.56%	58.82%
Year	100.00%	88.89%	94.12%
Surname	87.50%	77.78%	82.35%
GivenName	87.50%	77.78%	82.35%
Volume	100.00%	66.67%	80.00%
FirstPage	83.33%	71.43%	76.92%
LastPage	83.33%	71.43%	76.92%
OVERALL	87.50%	77.78%	82.35%

Table 4.1 Precision, Recall, F1 score values

The classification category is depicted in the Fig4.2 and the confusion matrix is generated from the observed results.

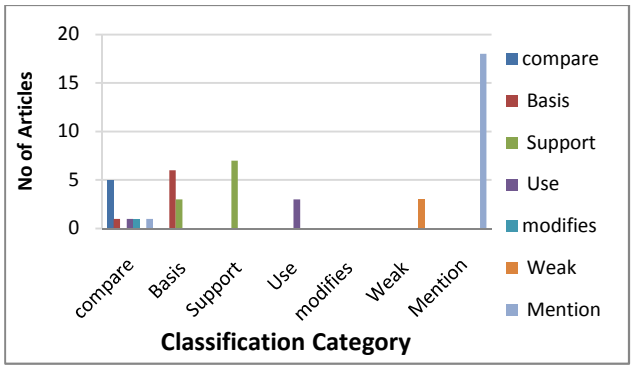

Fig 4.2 Confusion Matrix

As expected, the distribution is much skewed, with more than 60% of the citations of category *Mention*. The interesting phenomenon is the relatively high frequency of usage categories (Use, Modifies, Base, Compare contrast) with a total of 18.9%. There is a relatively low frequency of clearly negative citations (Weak total of 4.1%). The reason may be because the weak category is lower and mention category is higher. This may be because the author hesitates to refer other author with explicit negative sentences.

The Fig 4.3 depicts the sentiment identified as positive and negative accuracy using the SentiWordNet. In our experiments 80 cite sentences are examined of which 7 articles were identified as negative which are originally related to positive category. This is because our system does not consider the negation sentences.

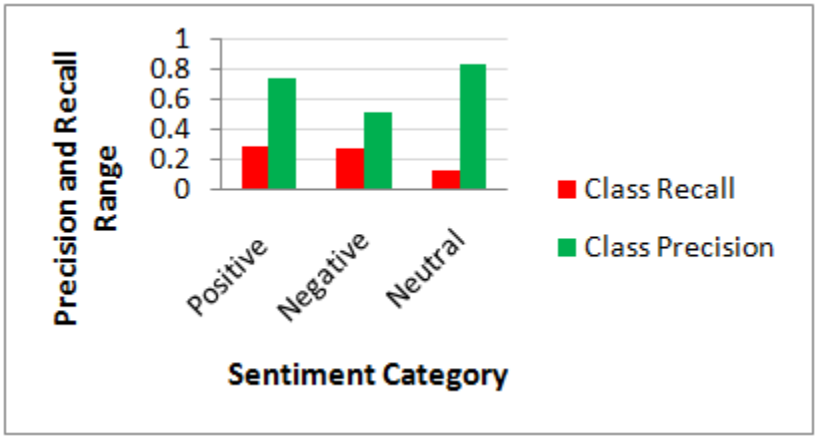

Fig 4.3 Sentiment **A**nalysis Score

The Table 4.2 depicts that this system identifies citation quality in terms of semantics and correctly predicts quality score The dataset for the experiment is 7 articles and their citation articles.

Articles	Cite Count	Citation Quality Score
Hwang,2010	9	0.266228934
Nakano,1980.	1	0.070298549
Raedt,2003	27	0.645416667
Fung,2004	33	0.508271324
Bruno, 2005	23	0.1984105
Kelleher,2006	23	0.373742086
Adler,2006.	35	0.362778596

Table 4.2 comparison between cite count and CQS

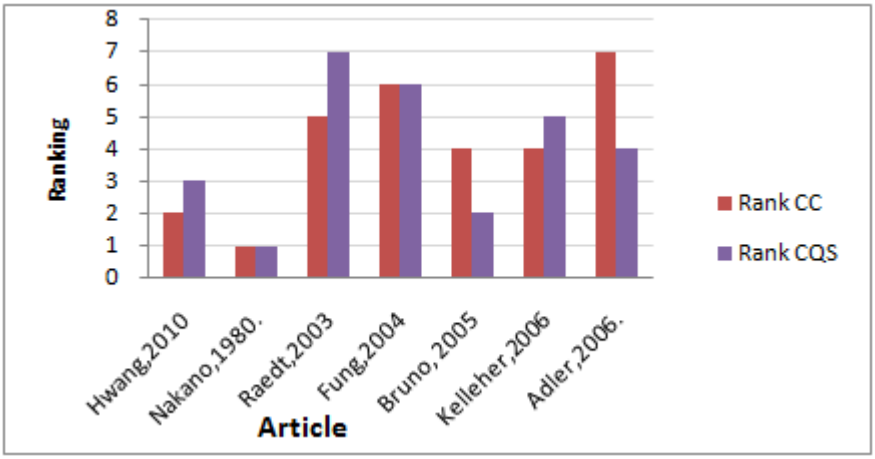

Fig 4.4 Ranking Score between cite count and CQS

The above diagram Fig 4.5 clearly depicts the difference between the citation count based rank and the citation quality score that we found. From the Table 4.2 it can be seen that Raedt, 2003 has the highest rank in terms of cite count whereas our system clearly predicts that Adler, 2006 holds highest rank in terms of semantic based evaluation among the 7 articles. Fung, 2004 is given the same rank in both the methods. From the above results it can be proved that citation count is not the only measure to determine the quality of article. It should include the semantics of the citation.

5. CONCLUSION

The system presented here identifies the citation quality for an article in the data in digital bibliographic repositories. We have provided a new method to rate a research paper based on its citation quality. We applied and implemented a supervised machine-learning system based on CRFs for citation parsing and report 80.05% F1-score to parse a citation into a total of eight fields. Our results show that CRFs are efficient machine learning model for citation parsing. SentiWordNet is valuable lexicon provides accurate values. The in-depth classification measure used in this work enables to exactly determine citer motivation.. All these techniques support exact calculation of citation quality. One issue we faced was that the number of citation

downloading is limited in our system because many articles do not have open access. This result generated by the system can be used in graph based work like citation network and find the main parts of research essence.

ACKNOWLEDGMENTS

This work was funded by Department of Science and Technology, India, under Fast Track Scheme for Young Scientists sanction no. SR/FTP?ETA-111/2010 dated 22.02.2012.

REFERENCES

[1] A.N. Langville and C.D. Meyer, "Google's PageRank and Beyond: The Science of Search En-gine Rankings", Princeton Univ. Press, 2006.
[2] AAAI-99 Workshop on Machine Learning for Information Extraction, pp. 37-42, 1999.
[3] Awais Athar, "Sentiment Analysis of Citations using Sentence Structure-Based Features", Proceedings of the ACL-HLT 2011 student session, portland, pp.81–87, june 2011.
[4] Diana C. Cavalcanti and Ricardo B. C. Prudêncio, "Good to be Bad? Distinguishing Between Positive and Negative Citations in Scientific Impact", 23rd IEEE International Conference on Tools with Artificial Intelligence, pp.156-162 2011.
[5] E .Garfield, "Journal impact factor: a brief review". CMAJ 1999, pp.979-980, 1999.
[6] E. Garfield, "Citation indexing: Its theory and application in science, technology, and humani-ties", John Wiley and Sons, Inc., New York, NY, USA, 1979.
[7] E. Garfield. "Citation analysis as a tool in journal evaluation". Science, Vol. 178, No. 60, pp.471–479, November 1972.
[8] Eugene Garifield. Can citation indexing be automated? In Mary Elizabeth Stevens, Vincent E. Giuliano, and Laurence B. Heilprin, editors, Statistical Association Methods for Mechanized Documentation, Vol.269 of National Bureau of Standards Miscellaneous Publication, Washington, Symposium Proceedings, pp.189-192, December 1965.
[9] G.S. Mahalakshmi, S. Sendhilkumar, and S. Dilip Sam, "Refining Research Citations. through Context Analysis", First International Symposium on Intelligent Informatics ISI 2012, Springer, 2012.
[10] Goodrum, K. McCain, S. Lawrence, and C. Giles, "Scholarly publishing in the Internet age: a citation analysis of computer science literature", Information Processing & Management, Vol. 37, pp. 661-675, 2001.
[11] H. Nanba, N. Kando, and M. Okumura. "Classification of research papers using citation links and citation types: Towards automatic review article generation", 11th SIG Classification Research Workshop, Classification for User Support and Learning, pp.117-134, 2000.
[12] Henry H. Bi, Jiamusi Wang, and Dennis K.J. L, "Comprehensive Citation Index for Research Networks", IEEE Transactions on Knowledge And Data Engineering, Vol. 23, No. 8, August 2011.
[13] Hidetsugu Nanba and Manabu Okumura. "Towards multi-paper summarization using reference information", Thomas Dean, editor, IJCAI, . Morgan Kaufmann, pp.926-931,1999.
[14] J. Deepika, G. S. Mahalakshmi, "Journal Impact Factor: A Measure of Quality or Popularity?", IICAI 2011, pp.1138-1157,2011.
[15] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment", Journal of ACM, Vol.46, No.5, pp.604–632, 1999.
[16] Jose A. de la Penaa, "Impact functions on the citation network of scientific articles", Journal of Informetrics, Elsevier, Vol. 5, No. 4, pp.565-573, 2011.
[17] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Or-der to the Web," technical report, Stanford Digital Library Technologies Project, 1998.
[18] Michael J. "Some results on the function and quality of citations" Social Studies of Science,Vol.5 No.1,pp.86-92, February 1975.

[19] O'connore, "Citing statements: computer recognition and use to improve retrieval", Information processing & management,Vol.18,No 3,pp 125-131,1982.

[20] Paul Zhang, Lavanya Koppaka ."Semantics based legal citation network",proceeding in ICAIL '07,pp.4-8, Palo Alto,USA ACM, June 2007.

[21] Qing Zhang, Yong-Gang Cao, Hong Yu "Parsing citations in biomedical articles using conditional random fields", Journal Computers in Biology and Medicine, Vol. 41 No. 4, pp 190-194,April 2011

[22] S. Lawrence, C. L. Giles, and K. Bollacker, "Digital libraries and autonomous citation indexing", Computer, vol.32, pp. 67-71, 1999.

[23] Shannon Bradshaw, "Reference directed indexing: Redeeming relevance for subject search in citation indexes", Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries, pp.499-510,2003.

[24] Simone Teufel, Advaith Siddharthan, and Dan Tidhar "An annotation scheme for citation function", 7th SIGdial Workshop on Discourse and Dialogue,Sydney, Assosiation for Computational Linguistics. Pp.80-87, July 2006.

[25] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. "Automatic classification of citation function", 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), Association for Computational Linguistics ,Sydney, pp. 103-110,July 2006.

[26] Son Bao Pham and Achim Hoffmann, "A New Approach for Scientific Citation Classification Using Cue Phrases" , proceeding of Australian joint conference in Articial Intelli-gence,Springer Vol. 2903, pp. 1759-799,2003

[27] Y. Ding, G. Chowdhury, and S. Foo, "Template mining for the extraction of citation from digital documents", Proceedings of the Second Asian Digital Library Conference, Taiwan, pp. 47-62. 1999.

[28] Z. Nie, Y. Zhang, J. Wen, and W. Ma, "Object-Level Ranking: Bringing Order to Web Objects", Proc. of the 14th Int'l World Wide Web Conf. WWW, pp.567-574,2005