RESEARCH ARTICLE

Editorial Process: Submission:01/13/2018 Acceptance:02/17/2019

Using Citation Context to Improve the Retrieval of Research Article from Cancer Research Journals

Parthasarathy G[1]*, L Lakshmanan[2], L Ramanathan[3]

Abstract

Objective: In recent years, citation analysis tools provide many devices for finding or computing the citation score or impact factor for journals. It is important for the researchers to identify good journals for collecting research ideas discussed. A journal with a good impact factor value is preferably referred to by many researchers. In this research work, the author proposes a system for ranking journals on the basis of ideas and results cited in other papers. **Methods:** The work involves the cited content extractor for extracting the descriptive features mentioned about the cited paper. The cited content refers to the content in the article written by a citing paper and relating to the cited paper. The ranking system uses a citation score estimator for computing the overall weight of the descriptive cited content relating to a specific paper in the citing papers. The journal ranking system performs classification of the citation content with the evaluation of a citation score. The work that involves the citation content is classified under different categories as positively cited, negatively cited or neutral and unrelated. **Results:** Then the computed citation score is used for ranking the dealing with research on cancer research journals. The results of the ranking journals indicate that the particular ranked journal has been cited in the literature of many journals with a good descriptive content. Journal ranking system can be considered as a well-organized tool for ranking the cancer research scientific journal based on citation content and citation counting. **Conclusion:** This experimental cancer journal ranking method increases accuracy and effectiveness by using the citation content when compared with PageRank and HITS.

Keywords: Opining mining- citation ranking- citation classification- cancer research journal- Information retrieval

Asian Pac J Cancer Prev, **20 (3)**, 951-960

Introduction

Extreme leverage of accessing online resources has led to a rapid increase in online access by researchers who do research by referring to various research ideas published in different journals. Most of the researcher's authors of articles write about cited papers in their research papers. The cited content may have opinions on the technique proposed in the cited paper. Cancer journal ranking system is useful for ranking and providing desirable recommendations of various well cited journals. Various citation analyses or journal ranking techniques are available and those systems might have used the citation count based ranking system. In addition, none of the systems is available for ranking journals based on the descriptive cited contents or the citation score calculated using the cited content. In respect of the computation of the citation score based on the cited content, there is a need for performing opinion mining on the cited content seen in a variety of research articles in literature.

Opinion mining can be referred to as a natural language processing of a type which involves natural language related tasks for finding out the users/authors' mindset about the particular journal which they have cited in their research papers. In general, opinion mining is the process of parsing the content, identifying the opinionated sentences and then finding the polarity of the opinionated sentences. The significant factor that accounts for the application of the opinion mining techniques in research papers is the desire to locate the level of opinions on a specific innovative technique seen in a journal. In addition, with the citation score computation, calculation of the opinion levels is of immense help, facilitating identification of the reputed journal or a novel idea which is the subject meter of discussion by different authors. The citing paper discusses or comments about the idea or results seen in the cited papers. Likewise, many different citing authors refer to or cite various researches works in their research papers. In such a scenario, it is important to identify the type and the level of opinions shared a particular research paper showed. Assessment of the reputation of the impact of a specific research journal cited in other journals is of

*[1]Department of Computer Science and Engineering, Jeppiaar Maamallan Engineering College, Anna University, [2]Department of Computer Science and Engineering˛, Sathyabama Institute of Science and Technology, Chennai, [3]School of Computer Science and Engineering (SCOPE),VIT, Vellore, India. *For Correspondence: amburgps@gmail.com*

immense help. Despite the availability of many citations analysis tools for providing the consolidated status about citations highlighting the reputation of a particular author or journal, it adds to the privilege to the citation if those cited papers are ranked on the basis of the opinion based score value.

The proposed ranking of a cancer research journal is an efficient ranking technique based on citations content and are classified under different classes as positively cited, negatively cited or neutral and unrelated. This ranking technique improves the accuracy through use of citation content when compared with the present one.

Materials and Methods

Research Methods

Citation analysis is the task of analysing the attributes of the citations that have been referred to in many journals by different authors. The attributes of the citation is collection of the author's name, journal name and the details of the publication. It also includes the process of performing analysis relating to the content of the particular citation cited in other papers. Opinion mining process becomes the need of the hour for a detailed process analysis. Opinions, sentiments and comments expressed online comprise the area of opinion mining, which very often involves the location and understanding of the content relating to a specific event or object. In the case of citation analysis, it is likely that the opinion relating to the research experimentation and its evaluated results is shared. Authors of research papers, as a common practice, have the liberty to present their views and comments on any paper seen in literature. Such opinions can be considered as useful tools to do the location and analysis of the factors that account for the popularity of the research papers. Currently the quality of every research article is judged by the number of citations made of it, with many citation measures introduced. The citation content, h-index or g-factor finds extensive use in the rating of the journal or as index of the researcher's popularity. Citation measurements have, as their core points, factors that include Elsevier's Scopus, Thomson Reuter's Web of science and Google scholar used as mining tools. But, citation mining methods do not unearth the entire related citations. Many concepts and ideas have been proposed with the objective of improving the quality of the citation measures. Some of these have made a significant contribution to authentic numeric evaluation (Birger, 2013). The task of organizing information is more important than citation analysis with ability to tag the cited content for identifying the opinion, which can be done effectively through the use a semantic tag. The use of semantic tag is considered as an effective way for tagging the content.

Categorization of the content is highly useful, with the use of semantically tagged words along with the features of the context of categorization on most occasions (Sérgio, 2012). Knowledge organization is the traditional field of information science. Bibliometrics and citation analysis techniques help organization of the knowledge relevant to the scientific journals. Citation analysis and bibliometric methods are considered as suitable approaches for organizing the citation databases (Birger, 2013). Though the citation count has a major role in measuring the importance of scientific journals, Citation analysis considers the usage of ranking algorithms for ranking scientific journals. The citation count provides extensive support along with PageRank algorithms for performing citation analysis in very large citation networks. These measures offer a wide support for showing the impact of the publications in different fields. The PageRank algorithm is used in the evaluation of the journal's influence in various fields (Nan et al., 2008). There exist a vast, number of citation databases from which identifying the impact of researchers is very crucial on most occasions, the impact is measured by calculating the total count of the citations. The citation count depends not only on publication impact but also on the presence or absence of influence of the researcher along with highly cited publications. Many different methods have been recommended in the citation analysis papers for improving the ways for counting the highly cited publications (Ludo et al., 2013).

Performance of experimental analysis of the various methodologies is essential prerequisite for computing citation indicators. The methodologies require the utilization of the citation counts in their entirety or fractions thereof. Citation indicators have a pronounced effect when fractionalized between the authoring countries. Calculation of citation indexes is based on the fractionalized citation counts rather than an entire citation count. This is due to, the citations being indexed. They can also be used for better ranking as the citations are invoked for the computation of the citation indication based on either the author or country-wise fractionalized count (Dag et al., 2012). Many research papers pose various questions and promote research analysis on e-government research areas and the manner of their innovation in other scientific areas. Many techniques are available for performing social network and citation analysis. But there is still need for a common method for use in these analyses. The dataset also plays a prominent role in performing these analyses (Nusa, 2009). Performance of citation analyses has been seen on social science journals published in Australia on the subject of determination of the differences between the data gathered from Scopus and the Web of Science. Comparison of data gathered with rankings assigned has been made by specialized groups for the journals with a research assessment model. The ranking of journals was made using measures that included extended publication impact factor, h-index and modified diffusion factor. The results showed the Scopus database yielding higher citations for most of the journals. The association between the assigned rankings by specialized group and the rankings derived from citation data has been rather small (Gaby H. et al., 2010). Text and graph mining algorithms used for performing full citation analysis involves the functionalities of performance of publication ranking. The citation contexts are extracted from full-text publications, with citation represented with a probability with predefined topics. The citation graph is constructed on the basis of the publication or citation topic distribution. Calculation of its score is done on the basis of the importance attributed to the publication. The results have shown the ability of the full text citation with publication content used along with PageRank algorithm to improve the citation analysis process (Xiaozhong et al., 2012). The citations of papers published in conferences and journals have been evaluated. Such citations help providing learning analytics to the users. The emerged learning analytics field blends various academic areas with different diversified methodologies with many scientific assumptions. Both the learning analytics and citation analysis have contributed a lot and targeted various research fields to provide support (Shane et al., 2014).

Performance of citation analysis is not confirmed to scientific journals or research papers. It also provides a way for performing analysis on collaborative computing areas, and the task of engaging multiple participants who may be either researchers or educators with the support of identifying journals with highest impact on the collaborative computing area (Clyde et al., 2003). Not all the conference papers or journals are always cited in many different journals by different authors. Only a very small bunch of papers are cited most frequently with a larger number of citations. Most of the cited journals belong to the education category and that too those that focus on computer science education and computing content. Many conference and journal publications are often measured with their citation impact level and even this is compared with different time periods of the publications (Raymond et al., 2010). Computation of the citation score of the journals can be done using different methods that included overall score, normalized score and weighted score. The scores of the journals are worked out and comparisons between the scores made. Some of these have ideal ranking in all the three methods. The comparative analysis has triggered many discourses on scholarly research, bringing the effective role of the journals to the surface (Chun Hung et al., 1999). Fuzzy FCA based approach is one of the techniques for conceptually clustering the uncertainty data. Clustering technique helps to generation of concept hierarchy of research areas from citation database. The clusters are generated and the relations are represented among them hierarchically (Petra et al., 2010).

The intellectuality of any research domain can be structured with author co-citation analysis being used for the identification of the intellectual structure. The focus is on the citation count without any involvement of the citation content. Similarity between the co-cited authors is measured on the consideration of the author's citation content. Full-text journal articles have been collected and the sentences cited have been extracted for computing the similarity distances (Yoo et al., 2014). Citation analysis has become an essential tool for the evaluation of the institutions. The citation impact has to be measured by evaluating the citation data. Any citation data that includes very recent data is considered data as meaningful for the evaluation of the publications (Lutz et al., 2013). The mathematical functions have found in the analysis of the citation distribution of research papers to the specific authors. Mathematical functions that find the largest use are power-law, logarithmic, binomial, stretched exponential and Langmuir type. Analysis of the citation distribution of papers is done using the value computed that uses these mathematical functions (Keshra et al., 2013). The importance of the journals can be assessed for improving the quality of journals in those disciplines. The generalized impact factor is used with a combination of historical citation observed and future citation of the journal. The impact factor is used for ranking the relative influence of many journals (Zhou et al., 2009).

Various opining mining tools are available for the assessment of the semantic orientation of documents, one of which uses a hybrid method. All semantic patterns follow the basic formation of natural language grammar structure. Lexicon approach uses sentimental characteristic words while the classification approach uses K-Nearest Neighbor or Support-Vector Machine algorithms (Hai-Bing et al., 2011). A consideration of the review analysis process shows the use of the natural language in most of the online data reviews, in an unformatted data form. Cited information also appears in a natural language format. The analyses are called upon for checking their sentimental level and utilizing the computation process to a large extent in view of to the absence of automation of the analysis. The reviews go through automatic imposition in the process of checking the relevant opinions that are characterized. Dependency analysis is done in respect of the syntactic features of the review information (Somprasertsri et al., 2010). Link based ranking algorithms utilize the potential power of bibliographical citations for information retrieval in large online libraries (Larsen et al., 2006). Page rank and HITS are widely used in most of the citation tools ranking the citations, since both ranking algorithms provide good results for ranking the citations in large citation databases (Su et al., 2009). This paper has utilized chi-square test and statistical analysis for text mining for classification of Google Japan and Yahoo! Japan (Tsuyoshi et al., 2017). Breast Cancer has been detected earlier in Mexico based on competency based methods and also implemented through (MOOC) (Laura, 2018).

The new methodology introduced to overcome the cancer research obstacles using a technique of cell reprogramming for cancer modeling, treatment and solution to the particular obstacles (Saito et al., 2019). This paper give future need of requirement in expert, individual, furthermore, development for clinicians and wellbeing experts is basic to enhance nature of cancer care and refreshed wellbeing correspondence with patients and relatives (Shankar et al., 2018).

System Architecture

System Architecture used in the ranking of medical journal is shown in Figure 1. It consists of various components, namely database repository of cited documents, sentence parser for parsing the cited document content, repository of cited contents, cited opinion extractor, cited document score estimator, repository of cited documents' score and entity based ranker.

Repository of Cited and Citing Papers

The database repository consists of numerous documents referred to as cited documents. The term cited document demonstrates that the particular research paper/document has been referred to in various other research papers. The citing paper is which just refers to many other papers relevant to its area of experiment like biomedical. Both these cited and citing papers are taken into consideration for performing citation analysis for identifying the paper which has been extensively cited in various papers.

Cited Content Parser and Cited Opinion Extractor

Content parser is the normal natural language processing parser for parsing the sentences with their noun, adjective, adverb and etc. A cited content parser involves the functionalities of parsing the contents of the cited document. It helps splitting of the cited document content into noun, verb, adjective, adverb and etc. Consider for example, the sentence "Compare the results between supervised and unsupervised techniques and conclude that supervised machine learning was "more reliable". This sentence can be represented using Stanford Dependency parser as follows:

```
root (root-0, compare-1)
det (results-3, the-2)
dobj (compare-1, results-3)
prep (results-3, between-4)
amod (techniques-8, supervised-5)
cc (supervised-5, and-6)
conj (supervised-5, unsupervised-7)
pobj (between-4, techniques-8)
cc (Compare-1, and-9)
conj (Compare-1, conclude-10)
mark (reliable-18, that-11)
amod (learning-14, supervised-12)
nn (learning-14, machine-13)
nsubj(reliable-18, learning-14)
cop (reliable-18, was-15)
advmod (reliable-18, more-17)
ccomp (conclude-10, reliable-18)
```

The parsed opinionated sentence contains different parts among which the adverb modifier is taken into consideration. It contains only the adjective with the adverb which has been the subject matter of comment in the citing papers about the cited paper. The opinion extractor has the functionalities of the sentence parser used in the natural language processing. It helps extraction of the cited content with the commented adjective and mostly promotes or demotes the quality of the work performed in any cited paper. The extracted adjective may be either positive or negative, depending upon the context, that is, whether the term is positive or negative and the score is computed later. The parsed contents are stored in a repository which contains the cited contents of all the documents. The repository helps retrieval of the parsed content for computing the weight of the opinion commented on that particular cited paper in the citing paper.

Cited Content Score Estimator

The score estimator involves the process of assessment of the weight of the adjective used in the citing paper and related to the research work of the cited paper. The score value ranges from 0 to 1. The bags of adjective terms with their corresponding scores are used for assigning the scores to the adjective terms used in the research paper. POS tagging is applied on the cited sentences for segregating the noun, adjective, adverb and other subjective, objective related content that are seen in the literature of any citing paper. Though the individual tagged terms may not exactly predict the purpose or the intention for which it has been commented, it is helpful in getting the knowledge of the features that have been commented as noun terms and the adjective term that have been used for the description of the quality of the idea discussed in a research paper. The authors have used parsing tagger for tagging the individual words that have ultimately been used for qualifying the content. The tagger gets the cited content as the input and then checks for the noun, adjective and adverb terms those are associated with the cited content. Those tagged words are used later for computing the weight of the cited content on the basis of the descriptive tagged words available in the cited content.

The descriptive contents about the cited paper have been tagged for getting the subjective and modifier terms. But this does not exactly predict the relationship that exists between the general and the opinionated words that have been commented upon in literature. Hence , the authors have used the Stanford sentence parser for parsing and getting the sentence pattern which the formation of the overall review content revealing the relationship between the corresponding terms that have been the subject of description. Generally the parser parses the sentence on the basis of the words that help getting into the exact descriptive content of the paper cited. The authors customized the implementation for getting the noun subjective, adverb modifiers and the negation terms that exist in the literature content. This is so, considering that as the important factors mainly involved in any review sentence for describing the product. Once those terms have been identified, they have to be invoked for the calculation of the weight of cited content based on similarity level that the users have used to describe the research paper.

Opinionated Citation Score Database

The opinion scores of all the described sentences together with the corresponding cited papers' citations are stored in the database for further processing that includes indexing and ranking of the cited documents. A single research paper might have been cited in many different research papers known as citing papers. The cited documents are considered as different entities on the basis of their corresponding scores. Then those entities are ranked on the basis of their score values already computed. The cited contents are considered as entities as the cited contents are stored together with their corresponding opinion. Various cited papers with different score values are ranked on the basis of their citing papers in the same manner.

The weight prediction process includes sentence parsing and assignment of values based on the parsed phrases. The process consists of a few cases for assigning weights to sentences and phrases. The following two cases are considered for computing the weight of simple adjective and adjective with adverb present in the reviews.

Case 1: Simple Adjective

0< Weight of Adjective < 0.5, if the adjective is negative (1)

0.5 < Weight of Adjective < 1, if the adjective is positive (2)

The adjectives are extracted using the Stanford type dependency parser and the extracted adjectives are invoked for computing the score. Score values are assigned for the adjectives on the basis of the formulae mentioned in (1) and (2). If the extracted adjective is positive then the value is randomly assigned is between 0.5 and 1. In the implementation stage, the authors have used the random number function. Similarly, If the extracted adjective is negative, a value is between 0 and 0.5 is assigned. The adjectives are the core words used in the cited content to describe the research paper. It is important to assign score values for the adjectives used in citing papers.

The scores for the adjective used in cited content has been shown in Table 1.

Positive Weight of Adjective with adverb = sqrt(adjval) if adjective is positive, 0.5< adjval< 1 (3)

Negative Weight of Adjective with adverb = pow (adjval,2) , if adjective is negative,0 < adjval<0.5 (4)

The cited content mostly consists of adjectives used either for a high praise or a bad denouncement of the specific paper. If the commented sentence has a positive adjective along with an adverb, then the score value is computed based on the formula represented in (3). The score value is the square root of the adjective term used. The numeric value used for square rooting is the value assigned by the random number function and it is in the range of 0.5 to 1. When the sentence has a negative adjective along with the adverb then the score value is computed based on the formula represented in (4). The score value is the power two of the adjective term. The numeric value used in the power two is the value assigned by the random number function and is in the range of 0 to 0.5. The value of the power two is always less than the number which is powered by two.

The citation score measures for the adjectives with adverbs have been presented in Table 2. which shows the adjectives used in the citing paper about the cited paper.

Score (term) = [pos_score]-[neg_score] (5)

For instance, a particular author has given a description of another research paper and then, adjective with adverbial words is extracted and the score is computed. The positive and negative terms contained in that particular sentence are taken into consideration and the score is computed using the formula mentioned in (5).

Table 2 showing the citation score measures for adjective with adverb then score is computed in the same manner for all the descriptive sentences described in other papers which have cited this particular paper P.

Citation-Score=1/n ∑Score (termi), where i=1 to n (6)

The citation score is the impact value of a particular paper P that has been cited in n papers. The computed opinion score of a particular paper P in all other papers (termi) is taken summation and it is considered as the citation score of that particular paper. Similarly the citation score is computed for other papers also. There are totally nine papers have been taken into consideration for experimentation. The citation scores have been computed for all those nine papers.

Proposed Medical Journal Ranking Algorithm

Given a domain keyword K as the context and the collection of all related papers P_1 to P_n are downloaded the collection of all citing papers C_1 to C_n as PDF document. Classify the cited papers Ci on the basis of citation score be it positive, negative or neutral and unrelated.

Algorithm Steps:

INPUT: Domain Keywords

OUTPUT: Classified Cited Content

Step1: Get input from the user as domain keyword "Breast Cancer" to the citation spider or bots.

Step2: Retrieve the related and cited papers from the web using the bots based on the domain keywords given.

Step3: Convert the extracted PDF documents to text using text converter tool PDFBOX available in

Open source Java.

Step4: Parse the cited contents of the cited document and store in the repositories.

Step5: Extract the essence of the cited paper content and initialize the threshold value for each citation

Content of paper Pi for i=1 to n based on formula given below.

Positive Weight of Adjective with adverb = sqrt(adjval)

If adjective is positive, 0.5< adjval< 1

Negative Weight of Adjective with adverb = pow (adjval, 2)

If adjective is negative, 0 < adjval<0.5

Step6: Evaluate the score for the cited content using the formula given below.

Citation-Score=1/n ∑Score (termi), where i=1 to n

Step7: Classify the citation content score as positive, negative, neutral and undefined with the use of

citation classifier.

Step8: Rank the journal based on the citation score.

Implementation

Configuration of the web crawler is one for recursive retrieval of the websites related to the given input seeds. The input keywords are the key terms used in the search for the research articles in the digital libraries such as IEEE explore, ACM digital libraries, Sciencedirect, Google scholar and etc. The key terms specify those used for representing the domains with some other combination of search words. For example, the keyword “Breast cancer” helps retrieval of the research papers relevant to the particular domain. Table 3 shows the keywords and the number of citations that have been collected.

Following the downloading of the documents, the PDFs need to be collected and converted into text. The authors have used the PDFBox engine from Apache for the conversion and created a script to enable performance of the task. The total number of downloaded papers was around 2,191,600. This may not seem a large amount. However, the size of the collection is about 3 GB for the papers alone. This probably explains the necessity for a careful crawl and downloading.

Results

The citation score is the impact value of a particular paper P that has been cited in n papers. The computed opinion score of a particular paper P in all other papers is taken as summation and considered as the citation score of that particular paper. The citation score is computed for other papers also in a similar manner. Nine papers in all have been considered for experimentation. Citation scores have been computed for all these nine papers. The authors have used a dataset consisting of papers from SSO Annual Cancer Symposium in all their experiments. The data were drawn from Google scholar digital library. The

Table 1. Citation Score Measure for Adjective Terms

Positive Adjective	Score Value	B	Negative Adjective	Score Value
Reliable	0.95	Fuzzy Measures of Simple Adjectives	Not Good	0.45
Effective	0.84		Utterly	0.49
Finely	0.67		Acceptable	0.35
Adopted	0.58		Not Great	0.25
Fantastic	0.88		Moderate	0.43

Table 2. Citation Score Measures for Adjective with Adverb

Positive Adjective with Adverb	Score	E	Negative Adjective with Adverb	Score
More reliable	0.95	Fuzzy Measures of Adjectives with Adverb	Unfortunately	0.45
More useful	0.84		Not a useful	0.49
Well adopted	0.67		Not supported	0.35
Relatively better	0.58		Unexpectedly	0.25
Good idea	0.88		Poor quality research	0.43

Table 3. Input Key Terms for Collecting Citations fFrom Digital Libraries

Keyword	Number of Papers Since 2014	Number of Citations
Breast cancer	364,000	20,516
Brain cancer	368,000	39,900
Kidney cancer	216,000	8,829
Liver cancer	319,000	16,821
Lung cancer	316,000	3,900
Pancreatic cancer	120,000	35,886
Skin cancer	343,000	6,641
Thyroid cancer	47,000	24,921
Ovarian cancer	98,600	11,456
Total		168,870

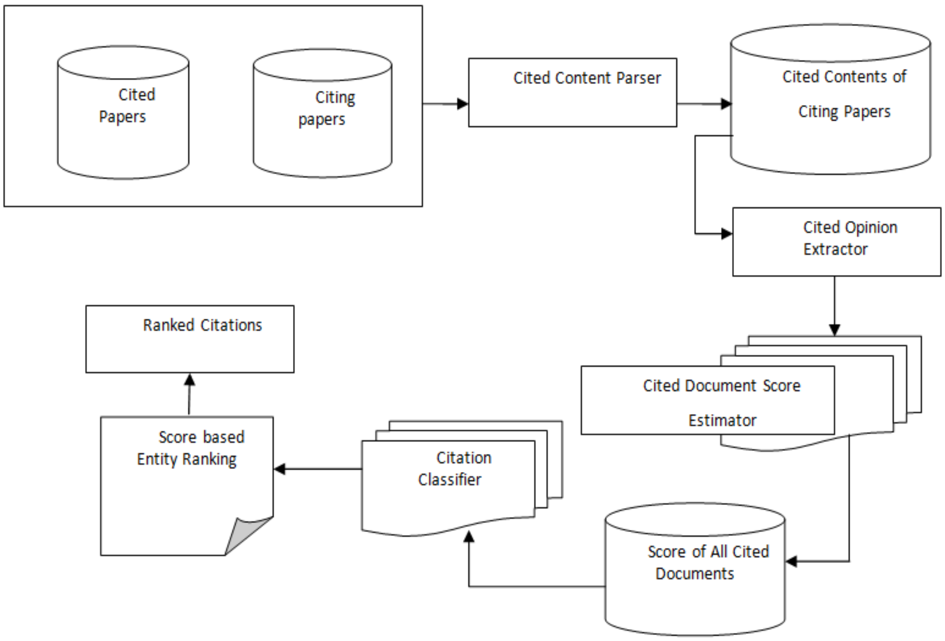
Cited Papers
Citing papers
Cited Content Parser
Cited Contents of Citing Papers
Cited Opinion Extractor
Cited Document Score Estimator
Score of All Cited Documents
Citation Classifier
Score based Entity Ranking
Ranked Citations

Figure 1. Medical Journal Ranking System Architecture

Table 4. Cited Papers and Its Citation Counts

Paper ID	Paper Title	Citation Count
P1	Recurrent and functional regulatory mutations in breast cancer, E Rheinbay. et al. (2017).	18
P2	Fluorescence navigation with indocyanine green for detecting sentinel lymph nodes in breast cancer, T Kitai. et al. (2005).	155
P3	Gene expression profiling predicts clinical outcome of breast cancer, LJ Van't Veer. et al. (2002).	9,091
P4	Prospective identification of tumorigenic breast cancer cells, R .Wooster. et al. (1995).	9,384
P5	Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer , DJ Slamon. et al. (1989).	7,379
P6	PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer , J Li. et al. (1997).	5,669
P7	Global burden of breast cancer , J Ferlay. et al. (2010).	2,081
P8	Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies , BD Lehmann . et al. (2011).	1,985
P9	The benefits and harms of breast cancer screening: an independent review, The Lancet (2012).	821
P10	Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer, A Prat. et al. (2010).	1,341
P11	The treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer, A Goldhirsch . et al. (2013).	1,516
P12	Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists , AC Wolff . et al. (2013).	1,891
P13	Randomised controlled trial of conservation therapy for breast cancer: 6-year analysis of the Scottish trial, AP Forrest et al. (1996).	440

dataset for the cited papers has been collected performing a journal paper search using the keyword “Breast cancer” and many papers have been listed out. For demonstration purposes, the authors taken the papers listed in Table 4 as cited papers with their citation counts.

The citation score computation for the cited paper “Recurrent and functional regulatory mutations in breast cancer” has been shown in Table 5.

The score computation for the cited paper with the cited content is computed as follows, “The experimental results have stated the approach is “More relevance “, in the Pan-cancer analysis of the entire genomes”. The adverb modifier, that is, the adjective with adverb is ‘more relevance’. According to case 2 (i.e. adjective with adverb), positive weight of adjective with adverb is as follows, Sqrt(adjval)= sqrt(0.72) .

[The adjval is the number generated by random numbers and it has been checked greater than 0.5 according to case 2 assumption] Sqrt(0.72) = 0.85. This is absolutely true according to the sample citation measures that have been shown in Table 2.

Table 3 shows the number citation count that has been given in Google Scholar for the respective paper.

The journal papers have been ranked on the basis of the descriptive content present in their literature of the particular citing paper. The computation of the citation

Table 5. Score Computation of Cited Papers

Paper ID	Citing paper	Cited Content	Cited Term	Score	Citation Score
P1	Pan-cancer analysis of whole genomes, PJ Campbell et.al.2017	Transcription factors and other proteins interact with enhancers, silencers, boundary elements, and overall chromatin structure for conferring cell-specific regulatory responses. Recent studies have revealed the greater relevance of this interplay in cancer.	More relevance	0.85	
	DNA damage response gene mutations and inherited susceptibility to breast cancer, t mantere,2017	Of late, large-scale DNA sequencing has helped the well systematic characterization of the full mutation repertoire in breast cancer, providing insights into the mutated cancer genes and mutational processes of the disease	Well systematic	0.62	
	A pan cancer analysis of promoter activity highlights the regulatory role of alternative transcription start sites and their association with noncoding mutations, D Demircioğlu et. al,2017	One of the key properties of cancer is larger increase in mutation rates that can affect not only gene products, but also gene regulation	Increase larger in mutation	0.82	
	Systematic Identification and Analysis of Cell-state-associated c is regulatory Elements Using Statistical Approaches, Y Yang – 2017	Aberrant c is regulatory elements in cancer are poorly characterized and understood	Poorly characterized	0.45	2.94

Table 6. Ranking of Cited Papers Based on Computed Cited Score

Paper ID	Citation Score	Citation Ranking
P1	2.94	9
P2	3.26	8
P3	8.75	1
P4	7.55	2
P5	6.5	4
P6	6.65	3
P7	2.65	10
P8	4.2	5
P9	3.5	6
P10	3.25	7
P11	2.25	11
P12	1.15	13
P13	1.25	12

score involves the process of the score assigned for the individual paper which has been cited in many research papers. The cited papers have been ranked on the basis of the computed citation score value. The ranked papers have been shown in Table 6. The consideration of the descriptive terms used in the citing paper by the ranked citations is obvious. Despite a particular paper having a small number of citation count, it has sometimes been ranked as high, in view of the rich cited content present in the citing paper.

Citation Classifiers

Citation classifiers constitute the standard data mining classifiers. Some of the classifiers that have focused use in the performance of the citation classification are J48, Conjunctive rule, AdaboostM1, Naive Bayes, Sequential Minimal Optimization, IBK instance based, Random Forest and Random Tree. Table 7 illustrates the comparison between the classification outputs of the citations computed for SSO Annual Cancer Symposium (SSOACS).

The authors have utilized the Society of Sugiacl Oncology (http://www.surgonc.org) Annoted Bibliography (SSO_AB) and https://archive.ics.uci.edu in their experiment. The SSO_AB has been kept up by the general public of surgical oncology(SSO) and assembled into 10 classifications, each with respect to cancer of a certain kind. The data were drawn from CiteSeer, a digital library of papers presented in conferences, symposiums and journals in medical research. CiteSeer does the work of collecting medical papers posted on the Internet through direct link to publishers, conference sites and journals. The parsing of these articles is then done for finding the citations and descriptive information seen in each paper. It has over 7, 00,000 indexed papers in its database.

Table 8 depicts the class-wise detailed accuracy of the classifiers with the polarity based classification. Good precision and recall rates in the evaluation of the citation data given in most of the classifiers.

Figure 2 is the graph plotted for the cited papers with their citation count. Most of the earlier citation based

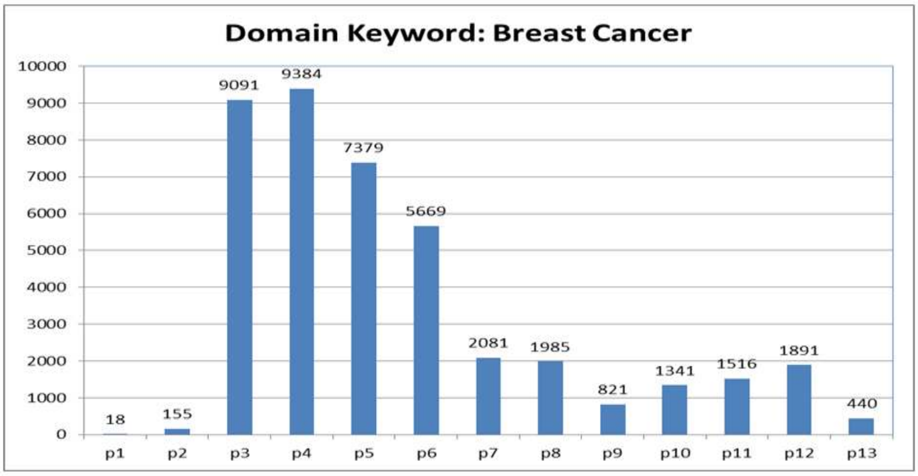

Figure 2. Citation Count Based Ranking between Cited Papers vs Citation Count

Table 7. Comparison of Different Classifiers' Results

Training Citations (33,774)/ Testing Citations (168,870)	Total Citations	Positive Citations	Negative Citations	Neutral Citations	Undefined Citations	Accuracy
SSOACS	168,870	126,270	1,350	5,040	36,210	74.78
SSOACS	168,870	121,620	1,410	4,920	40,920	72.03
SSOACS	168,870	121,260	1,380	5,040	41,190	71.82
SSOACS	168,870	120,990	1,380	3,510	42,990	71.65
SSOACS	168,870	118,770	1,170	5,010	43,920	70.33
SSOACS	168,870	117,090	1,440	3,960	46,380	69.34
SSOACS	168,870	116,490	1,440	5,040	45,900	68.11
SSOACS	168,870	114,180	1,440	3,780	49,470	67.33

Table 8. Class-Wise Detailed Accuracy

Classifiers	Detailed Accuracy	Positive	Negative	Undefined	Both
J48	Precision	0.767	0	0.322	0
	Recall	0.933	0	0.423	0
	F-measure	0.842	0	0.643	0
Conjunctive Rule	Precision	0.736	0	0.51	0
	Recall	0.954	0	0.145	0
	F-measure	0.831	0	0.226	0
AdaBoostM1	Precision	0.718	0	0	0
	Recall	1	0	0	0
	F-measure	0.836	0	0	0
Naïve Bayes	Precision	0.718	0	0	0
	Recall	1	0	0	0
	F-measure	0.836	0	0	0
SMO	Precision	0.718	0	0	0
	Recall	1	0	0	0
	F-measure	0.836	0	0	0
IBK Instance based	Precision	0.773	0	0.536	0.04
	Recall	0.826	0	0.407	0.07
	F-measure	0.799	0	0.463	0.06
Random Forest	Precision	0.785	0	0.474	0.05
	Recall	0.789	0	0.467	0.05
	F-measure	0.787	0	0.471	0.05
Random Tree	Precision	0.785	0	0.469	0.04
	Recall	0.782	0	0.469	0.04
	F-measure	0.783	0	0.469	0.06

journal ranking systems have considered only the citation count. In addition, it is to be noted that most of the impact factor methods make use this citation count data as the prime resource.

Figure 3 is a list the cited papers with the computed citation score. The journal papers that have been ranked on the basis of the score value computed using the proposed approach. A comparison with the earlier citation count based graph shows, this score based ranking providing additional support since those papers have been invoked to involve opinion mined content for score computation. We classify the papers from other papers in the above graph that paper p3 has accuracy higher than the other papers.

Discussion

In this work, a citation ranking system has been presented for the calculation of the weight of the opinion strength of the descriptive content in the citing paper. The descriptive opinionated cited content has been extracted

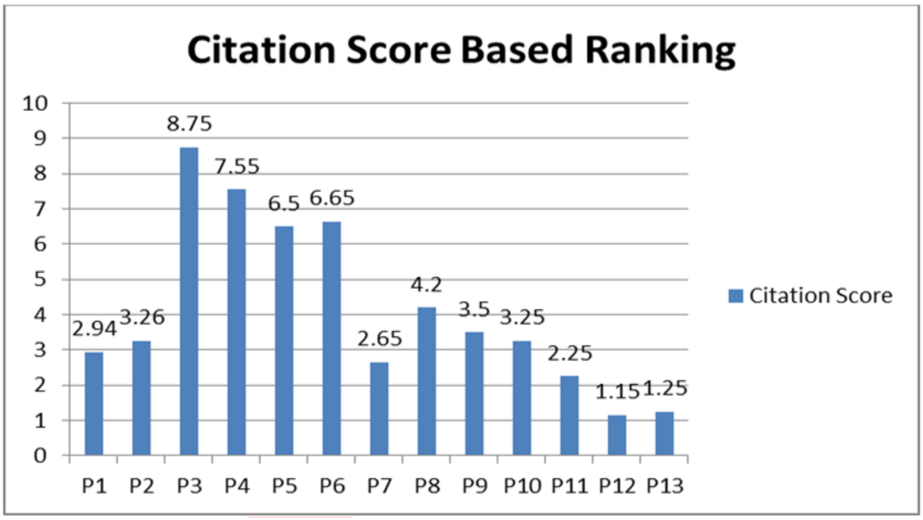

Figure 3. Cited Papers vs Citation Score

using the Stanford dependency parser. The extracted words were imposed for the application of the score computation process. The scores were computed on the basis of the available descriptive features, namely, the adjectives and adjectives with adverbs. Following the computation of citation scores, the journal papers are ranked on the basis of the computed citation score along with the use of particular paper's citation count. Currently, the classifiers make use of the words in the cited content for the training. Though the words present in the cited content provide the valuable information, still some meaningful information can be found in and around the cited content. Overlapping of cited content with different citations may be seen when the context is increased. In such cases, the learning model has to be adopted for handling the cited contents of two different citation contexts. In order to improve the accuracy of the classification, the large corpus can be built for cited content of the citations. In addition, the learning can be bootstrapped with the words used in rule based classifiers. Even the time information can also be used for considerably increasing the accuracy of the citation classification.

In earlier research works, the citation score or impact factor was calculated on the basis of the citation count alone. Then research works started using PageRank algorithm for ranking the citations. The PageRank algorithm assigns weights to papers in prorate to the importance of the paper. Modified PageRank algorithms have been used for improving the ranking process. It is also to be noted that none of the earlier works has indicated both the citation count and the computed citation score. The experimental outputs indicate the proposed technique as effective in predicting the popularity of the research papers. Further, the weight prediction method can be improved by checking the exactly relevant descriptive content by invoking the fuzzy-based content's score estimation. In this paper, the authors have provided a novel approach and prototype model called journal ranking system for citation ranking and selection of cancer research scientific journal by positive citation content using citation score measurement. As future work, the authors suggest the use of some more features when computing the citation score for locating the similarities between the citation content using clustering to improve efficiency.

References

Birger H (2013). Citation analysis: A social and dynamic approach to knowledge organization. *Inf Process Manag*, **49**, 1313- 25.

Chun Hung C, Kumar A, Motwani JG, et al (1999). A citation analysis of the technology innovation management journals. *IEEE T Eng Manage*, **46**, 4-33.

Clyde W, Holsapple D, Wenhong L (2003). A citation analysis of influences on collaborative computing research. *Comput Supp Coop W J*, **12**, 351-66.

Dag WA, Schneider JW, Gunnarsson M (2012). Ranking national research systems by citation indicators. A comparative analysis using whole and fractionalized counting methods. *J Informetr*, **6**, 36-43.

Gaby H, Genoni P (2010). Citation analysis and peer ranking of Australian social science journals, Scientometrics, Springer, **85**, pp 471-87.

Hai-Bing M, Geng YB, Qiu JR (2011). Analysis of three methods for web-based opinion mining. IEEE/ICMLC, pp 915-9.

Keshra S (2013). Comparison of different mathematical functions for the analysis of citation distribution of papers of individual authors. *J Informetr*, **7**, 36-49.

Larsen B, Ingwersen P (2006). Using citations for ranking in digital libraries. Digital Libraries, ACM/IEEE-CS Joint Conference, pp 370.

Ludo W, Van Eck NJ, Wouters P (2013). Counting publications and citations: Is more always better?. *J Informetr*, **7**, 635-41.

Lutz B (2013). The problem of citation impact assessments for recent publication years in institutional evaluations. *J Informetr*, **7**, 722-9.

Laura MV (2018). Training primary health professionals in breast cancer prevention: Evidence and experience from Maxico. *J Cancer Edu*, **33**, 160-6.

Nan M, Guan J, Zhao Y (2008). Bringing pagerank to the citation analysis. *Inf Process Manag*, **44**, 800-10.

Nusa E (2009). Citation analysis for e-government research, Proceedings of the 10th Annual International Conference on Digital Government Research: Making Connections between Citizens, Data and Government, pp 244-53.

Petra P (2010). Fuzzy conceptual clustering. 10th industrial conference on Advances in data mining: applications and theoretical aspects. Springer-Verlag, Berlin, Heidelberg, pp 71-85.

Shane D, Dragan G, George S, Srecko J (2014). Current state and future trends: a citation network analysis of the learning analytics field, Fourth International Conference on Learning Analytics and Knowledge, Pp 231-40

Shankar A, Thakur R, Meshram N, Keditsu K, Srinivas P (2018). NCI summer curriculum in cancer control and prevention– A practice changing course for oncologists from limited resource country like India. *Asian Pac J Cancer Prev*, **19**, 1157-60 .

Somprasertsri G, Lalitrojwong P (2010). Extracting product features and opinions from product reviews using dependency analysis. IEEE/ ICFSKD, pp 2358-62.

Su C, Pan U, Yuan J, et al (2009). Pagerank, HITS and impact factor for journal ranking. *J Comput Inf Sci Eng*, **6**, 285-29.

Saito S, Lin YC, Nakamura Y, et al (2019). Potential application of cell reprogramming techniques for cancer research. *Cell Mol Life Sci*, **76**, 45-65.

Tsuyoshi O, Hirono I, Masahumi O, Mio K, Takahiro K (2017). Assertions of Japanese websites for and against cancer screening: a text mining analysis. *Asian Pac J Cancer Prev*, **18**, 1069-75.

Xiaozhong L, Jinsong Z, Chun g (2012). Full-text citation analysis: enhancing bibliometric and scientific publication ranking. ACM international conference on Information and knowledge management, pp 1975-9.

Yoo KJ, Min S, Ying D (2014). Content-based author co-citation analysis. *J Informetr*, **8**, 197-211.

Zhou X, Yang G (2009). Using extended R-impact to assess journal influence. *IEEE T Reliab*, **58**, 317-23.