ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

RECOGNIZING SELF-CITATIONS VIA CITATION QUALITY ANALYSIS

[1]M. RAJA AND [2]T. RAVICHANDRAN

[1] Department of Computer Science & Engg., CMS College Engineering and Technology, Coimbatore, India
[2]Principal of Hindustan Institute of Technology, Coimbatore,Tamilnadu, India
Email: [1]rajajuly2014@gmail.com

ABSTRACT

Self citations have so far been excluded from citation count. It is widely believed that such self citations do not have any significance than merely to increase the citation count of the article and improve the prestige of the author through citation count.

Self citations should not always be avoided as the article may have been cited for genuine reasons. Analysis of self-citations helps in evaluating continuing research as well.

This paper argues that self-citations must not be blindly excluded in citation counts when evaluating prestige value of a research paper/author. Experiments conducted on researcher's self-citation dataset reveal that most self-citations show marginal improvement thus establishing researcher progress in the respective area of research.

Keywords: *Citations, Self Citations, Impact Factor, Citation Index, h-index*

1. INTRODUCTION

Citation is a reference to a published or unpublished source. More precisely, a citation is an abbreviated alphanumeric expression embedded in the body of an intellectual work that denotes an entry in the bibliographic references section for the purpose of acknowledging the relevance of the works of others to the topic of discussion at the spot where the citation appears.

Citations play a major role in representing semantic content of the full text articles. Citations help to document the source works that underpins a particular concepts, propositions and arguments. The purpose of these citations is to help readers identify and relocate the source work, provide evidence of research in the topic of concern and to acknowledge the findings of the author who originally contributed the concept or theory. Author self-citation contributes [Deepika and Mahalakshmi, 2011a; Deepika and Mahalakshmi, 2012] to the overall citation count of an article and the impact factor of the journal in which it appears.

Self-citing is often considered as an attempt of self-advertising. But an author self cites because the nature of work may build upon previous findings. *"Given the cumulative nature of the production of new knowledge, self-citations constitute a natural part of the communication process."* [Costas et al., 2010]. By including self-citation a certain problematic effect called PIED-PIPER effect can be eliminated. PIED-PIPER effect can be explained by considering a situation where in a low cited but an important paper of a researcher doesn't acquire the proper recognition due to blind elimination of self-citations can affect the researcher.

Self-citations are believed to be plagiarized across the parental context and are not given the worth they deserve. Journal impact factors do not measure the author quality [Deepika and Mahalakshmi, 2011a; Deepika and Mahalakshmi, 2012] whereas h-index has an inclusive measurement of self-citations. Again, it is policy wise and depends on the organization that counts citations for h-index analysis. Google Scholar includes self-citations into the total citations measure and Scopus exclude the self-citations for calculation of h-index of a particular researcher.

Analysing the quality of self-citations helps to build the prestige of researcher thereby improving the researcher presence in research communities [Mahalakshmi et al., 2012a & 2012b; Sendhilkumar et al., 2012]. Validity of self-citations needs to be explored since every self-citation is an indicative of continuing research [Mahalakshmi and Sendhilkumar, 2013].

This paper discusses a qualitative analysis of analyzing self-cited research publications from the perspective of citation quality. The research papers and their self-citations are tracked [Mahalakshmi and Sendhilkumar, 2008; Mahalakshmi et al., 2009; Mahalakshmi et al., 2011] and fed as input to the citation quality analyser. The (self) citations are analysed for citation quality via citation sentiment analysis [Sendhilkumar and Mahalakshmi, 2011; Mahalakshmi et al., 2013; Sendhilkumar et al., 2013c] which ranks the self-citations qualitatively. Self-citations beyond a definite threshold are rejected. Thus every researcher is recommended with quality self-citations which shall be included prestigiously in total citations thereby genuinely improving the h-index of the researcher.

2. LITERATURE SURVEY

Self citations had always been the topic of argument since the early research on citation metrics. Self Citations are not considered for the bibliometric evaluation of articles. It is widely believed that these citations do not necessarily reflect the importance of their work or its impact on the rest of the scientific community [James & Dag, 2007]. Early research on this was done on National Citation Report (NCR) for Norway ISI. It was concluded that self-citations have impact on the article. As the number of self citations increase the probability of the article being cited by others also increases. While it is already well-known that citation counts are at best a noisy indicator of scientific contribution, this work suggests an even deeper problem – even counts of citations from others are sensitive to strategic manipulation by those who are willing to cite themselves frequently.

The citation count and author's work is highly influenced by the number of self citations and recitations of the article [Isola et al, 2010]. A study reports that the scientometric measures like JIF [Deepika and Mahalakshmi, 2011b], h–index [Hirsch, 2005] etc are greatly affected by the self citations and the pattern in which they are cited [Lievers, 2012]. The forward-chronological reference resulted in large pool of potential articles. The articles were prioritized based on relevance among citations. The results showed that the f-N relationship was affected by self-citations. The study concluded that self references are more likely to be repeatedly cited and the repeated citation of self-references can be an indicator reflecting the true relevance to the citing document.

A study on when to begin citation counting disregards the previous measures where the articles availability played a major role in citation count [Craig et al., 2007]. This study argues that online availability of the article should not be affecting the citation metric. This also includes the time at which the article was published plays a major role for citation index.

Citations were analyzed and evaluated using various methods. Initially, frequency of citations received from Science Citation Index (SCI) database 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 introduction of graph-theoretic approaches, researchers were motivated to rank network entities using 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) [Henry H. Bi et al., 2011] considers both direct and indirect influence of research article. The indirect influence is considered by citation links with even those papers that do not directly cite it. Heterogeneous PageRank algorithm [Lagville et al, 2006] 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 negation. 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 [Simon Teufel et al., 2006].

A completely diverse yet simple classification was proposed which composed of only three categories namely Type B (base), C (compare) and O (other than B and C) [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.

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 following methodologies are proposed in our system for handling self-citations.

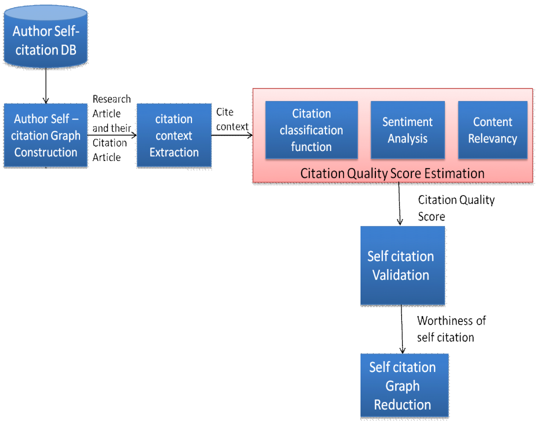
Author Self-citation DB
Author Self – citation Graph Construction
Research Article and their Citation Article
citation context Extraction
Cite context
Citation classification function
Sentiment Analysis
Content Relevancy
Citation Quality Score Estimation
Citation Quality Score
Self citation Validation
Worthiness of self citation
Self citation Graph Reduction

Figure 1: Overall Framework for analyzing citation quality of Self-citations

3.1 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.

3.2 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. This paper uses SentiWordNet [Baccianella et al., 2010] 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.

3.3 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, 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.

3.4 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 focuses 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 [Deepika et al., 2011; Deepika and Mahalakshmi, 2011b; 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.

3.5 Aggregating the Citation Score

In each citation article the seed paper may have be referenced two or more within the paper. Each reference point is called the citation instances. Classification is non numerical value that can be used for the purpose of combining cite instance values. Aggregation of quality score is based on the aspects of the citation. Citation quality score is calculated by the scores obtained from sentiment analysis, similarity and outlier detection. The classification categories are ordered based on the importance as: Compare 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.

$$Citation\ Quality\ Score = sentiment_i + Similarity_i + LDA_i \quad (1)$$

where i =highest importance classification category. In case of the Citation instance being an outlier then aggregation process omits the LDA similarity score.

$$Citation\ Quality\ Score = sentiment_i + Similarity_i \quad (2)$$

where i =highest importance classification category.

3.6 Self Citation Graph Reduction

Self citation network contain nodes and edges. Nodes represent the article. Edges represent the author self citation relationship between the articles. The graph representation for author Ying Dar Lin and his self-citations (as in Table 1) is given in Figure 2.

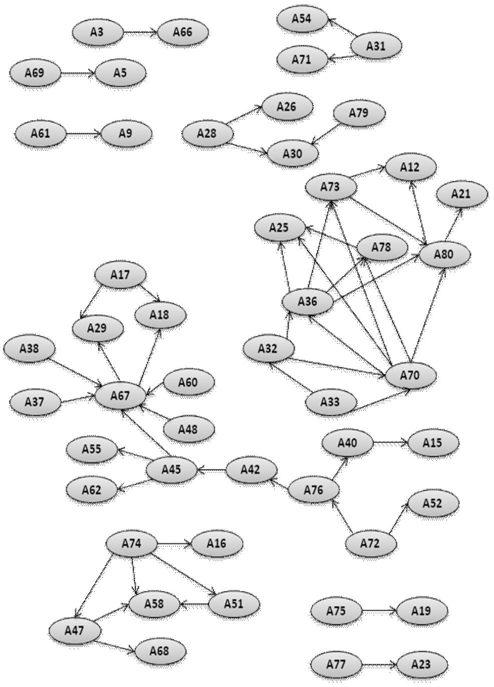

Figure 2: Self Citation network of Author Ying Dar Lin

The self citation network (figure 2) is reduced by considering the citation quality score across articles. This filtering will enable us to retain only the self citations that are semantically significant. The edges and the nodes that have less worthy citation to the cited article are removed from the graph. The articles retained after the filtration will have valid citations with significant improvement from the cited article.

4. RESULTS

4.1 Citation Context Extraction

The following table shows the recall, precision, and F1-score for the context extraction. 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.

Table 1: Precision, Recall, F1 score values

Fields	Total	Identified	correct	Precision	Recall	F1
Title	53	52	52	1.00	0.98	0.99
Source	53	52	47	0.90	0.89	0.90
Year	53	52	51	0.98	0.96	0.97
Surname	53	52	50	0.96	0.94	0.95
GivenName	53	52	52	1.00	0.98	0.99
Volume	53	45	43	0.96	0.81	0.88
FirstPage	51	51	50	0.98	0.98	0.98
LastPage	51	51	50	0.98	0.98	0.98
Overall	53	52	51	0.98	0.96	0.97

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

4.2 Citation Classification

The confusion matrix showing the classification category is depicted in the following graph.

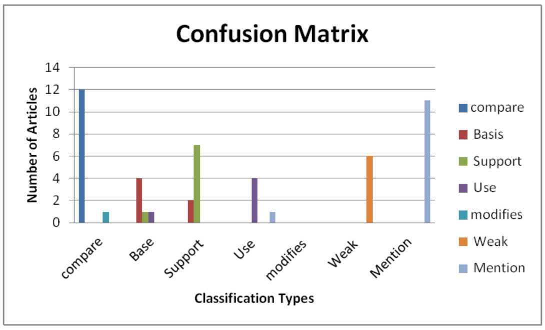

Figure 3: Confusion Matrix

From the analysis (figure 3), we saw that most citation was categorized as Compare. This is because in self citation the work is initially compared with previous work and the evolution of new work follows it. The second category is as expected mention category. Such articles can be exempted from citation count but the author working in the same domain cannot move to the next level without referring the problems related to the current issue. Such issues were evaluated in the content analysis part using semantic similarity and outlier detection techniques. The third highly classified category is Support. This pertains to the reality that most related articles of same author supports the cited articles. Very less articles were categorized as Base, Use, Modifies and Weak.

4.3 Sentiment Analysis:

The following graph shows the precision and recall of the positive and negative sentiment using SentiWordNet. A total of 52 cite sentences were examined, of which 5 citations were found to negative sentiment, 76% of sentences showed positive sentiment. This is because the author does not criticize themselves using explicit negative sentences.

Table 2: Sentiment Analysis Score

Class	Positive	Negative	Neutral
Predict Positive	40	1	0
Predict Negative	1	5	0
Predict Neutral	3	0	1
Total	52	52	52
Class Recall	0.769231	0.28	0.019231
Class Precision	0.741935	0.518519	0.833333

4.4 Self Citation Quality Score Estimation

Table 3: SCQS Score

S. No.	Article Id	SCQS	S. No.	Article Id	SCQS
1	A3	0.00452	15	A60	0.22679
2	A17	0.47439	16	A61	0.28433
3	A28	0.57558	17	A67	0.22432
4	A31	0.67213	18	A69	0.62131
5	A32	0.52421	19	A70	0.10127
6	A36	0.14895	20	A72	0.43562
7	A37	0.58307	21	A73	0.27657
8	A38	0.35917	22	A74	0.23021
9	A40	0.51925	23	A75	0.02413
10	A42	0.41932	24	A76	0.37858
11	A45	0.08731	25	A77	0.67655

12	A47	0.77801	26	A78	0.59955
13	A48	0.27894	27	A79	0.50413
14	A51	0.52115	28	A80	0.12439

Table 3 shows citation quality in terms of semantic citation quality score. Nearly 50 % of self citation quality is below average.

4.5 Graph Reduction

Figure 4 is the graphical representation of reduced author self-citation network. The directed edge starts at the cited article and ends at the citing article. The edge carries the citation quality score as weight.

Table 4: Article and self citation articles with SCQS

SI.No	Article Id	Citing Article ID	SCQS
1	A17	A18	0.2981937
2	A17	A29	0.6506
3	A28	A26	0.5458433
4	A28	A30	0.6053341
5	A3	A66	0.004526
6	A31	A54	0.5404161
7	A31	A71	0.8038475
8	A32	A36	0.5684659
9	A32	A70	0.4799559
10	A33	A32	0.5612969
11	A36	A25	0.7638528
12	A36	A73	0.488958
13	A36	A78	0.5593643
14	A36	A80	0.5958332
15	A37	A67	0.5830738
16	A38	A67	0.3591753
17	A40	A15	0.5192548
18	A42	A45	0.4193213
19	A45	A55	0.200565
20	A45	A62	0.200565
21	A45	A67	0.2619529
22	A47	A58	0.8419364
23	A47	A68	0.7140952
24	A48	A67	0.2789418
25	A51	A58	0.5211516
26	A60	A67	0.2267947
27	A61	A9	0.2843367
28	A67	A18	0.191106
29	A67	A29	0.4486462
30	A69	A5	0.6213123
31	A70	A25	0.6960017
32	A70	A36	0.7808646
33	A70	A73	0.5056398
34	A70	A78	0.5039071
35	A70	A80	0.5063536
36	A72	A52	0.4164363
37	A72	A76	0.4548092
38	A73	A12	0.2330252
39	A73	A80	0.3201237
40	A74	A16	0.2365377
41	A74	A47	0.4718693
42	A74	A51	0.5346419
43	A74	A58	0.9208537
44	A75	A19	0.0241305
45	A76	A40	0.2981866
46	A76	A42	0.4589841
47	A77	A23	0.6765528
48	A78	A25	0.5995552
49	A79	A30	0.5041364
50	A80	A12	0.2211541
51	A80	A21	0.0276352

The network is initially reduced using threshold on CQS score. The threshold is set as dynamic by considering the average of all CQS scores obtained. The edges having CQS below the threshold is deleted.

Table 5: Reduced Author Self Citations based on SCQS(refer Table 4)

SI.No	Article Id	Citing Article ID
1	A17	A29
2	A28	A26
3	A28	A30
4	A31	A54
5	A31	A71

6	A32	A36
7	A32	A70
8	A33	A32
9	A36	A25
10	A36	A73
11	A36	A78
12	A36	A80
13	A37	A67
14	A40	A15
15	A47	A58
16	A47	A68
17	A51	A58
18	A69	A5
19	A70	A25
20	A70	A36
21	A70	A73
22	A70	A78
23	A70	A80
24	A74	A47
25	A74	A51
26	A74	A58
27	A77	A23
28	A78	A25
29	A79	A30

The reduced graph figure 4 shows certain unconnected nodes. The reduction resulted in removing edges that had CQS below threshold. The unconnected edges denote the citations which has less semantic similarity and citation effect of the cited article. Such articles are to be avoided while calculating the number of citations. The nodes in which edges are retained must be included in the citation count as they have quality citation to the cited article.

The nodes that are disjoint do not have quality citations nor do they possess significant improvements, hence can be ignored in citation count. However, few discrepancies can be avoided by measuring novelty [Sendhilkumar et al., 2013a & 2013b] of research papers. We thereby put forth a strong argument that self-citations need to be analysed qualitatively and should definitely be honored by including them in citation counts.

5. CONCLUSION

Though bibliometrics research has failed to include self citations as a measure for citation index, they play a vital role in quantitative side. This paper has well analysed the research problem and recommended the self-citations based on citation quality. However, inclusion of self-citation is still a double-edged sword in bibliometrics research. Including self-citations will not cause a negative impact on the citation count as long as the paper address the problem in previous publication (self-cited article) from a different approach, identifies a new problem from the article or proposes a new methodology for the problem solved in previous publication with better results.

It can be inferred that most of the self citations shows very little improvement in work compared to the cited reference. Blindly removing such citations will not be convincing. Combining novelty and citation quality will help to decide whether or not to include the self citations and which citations are to be retained for further evaluations.

REFRENCES:

[1] A.N. Langville and C.D. Meyer, “Google’s PageRank and Beyond: The Science of Search Engine Rankings”, Princeton Univ. Press, 2006, AAAI-99 Workshop on Machine Learning for Information Extraction, pp. 37-42, 1999.

[2] 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.

[3] Costas, R., van Leeuwen, T.N., & Bordons, M. (2010). Self-citations at the meso and individual levels: effects of different calculation methods *Scientometrics* (82), 517-537.

[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] Deepika J. and Mahalakhsmi G.S., Journal Impact Factor – A measure of quality or popularity? – in proc. of int. conf. IICAI spl session on Advances in Web Intelligence and Data Mining – December 2011 (2011a)

[6] Deepika J. and Mahalakhsmi G.S., Comparative Evaluation of Text similarity Detection in Research Publications – in proc.

of int. conf. IICAI spl session on Advances in Natural Language Computing – December 2011 (2011 b)

[7] Deepika J., Archana V, Bagyalakshmi V, Preethi P and Mahalakshmi G.S., A Knowledge Based Approach to Detection of Idea Plagiarism in Online Research Publications, International Journal on Internet and Distributed Computing Systems, Vol. 1, No. 2, 2011, pp. 51-61.

[8] Deepika J. and Mahalakshmi G.S., Towards Knowledge based Impact Metrics for Open Source Research Publications, International Journal on Internet and Distributed Computing Systems, Vol. 2, No. 1, 2012, pp. 102-108.

[9] 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.

[10] E .Garfield, “Journal impact factor: a brief review”. CMAJ 1999, pp.979-980, 1999.

[11] G.S. Mahalakshmi and S. Sendhilkumar, Automatic Reference Tracking System, Handbook of Research on Text and Web Mining Technologies, edited by Min Song and Yi-Fang Wu, Idea Group Inc., USA, Vol. II, Section IV, Chapter XXIX, pp. 483-499, 2008

[12] G. S. Mahalakshmi, S. Sendhilkumar, Alagu Irulappan, Preetham Mirinda (2009), Ontology Based Relevance Analysis for Automatic Reference Tracking, International Journal of Computer Applications in Technology (IJCAT): Special Issue on Computer Applications in Knowledge-Based Systems, ISSN 0952-8091, Vol. 35, Nos. 2/3/4, pp. 165-173.

[13] G.S. Mahalakshmi, S. Sendhilkumar, Alagu Irulappan, Preethan Mirinda, and Gnanasekaran, Comparative Evaluation Of Ontology-Based Automatic Reference Tracking, International Journal of Networking and Virtual Organizations (IJNVO): Special issue on Open source Intelligence and Web Mining, Inderscience Publishers, Vol. 8, No. 1 / 2, 2011, pp. 142-157.

[14] G.S. Mahalakshmi, Dilip Sam S. and Sendhilkumar S., Establishing Knowledge Networks via Analysis of Research Abstracts, Special Issue of Journal of Universal Computer Science (JUCS) on “Advances on Social Network Applications”, Vol. 18, No. 8 (2012), pp. 993-1021.

[15] G.S. Mahalakshmi, Dilip Sam S. and Sendhilkumar S., Mining Research Abstracts for Exploration of Research Communities, ACM Compute 2012. – December 2012

[16] G.S. Mahalakshmi, S. Sendhilkumar, and S. Dilip Sam, “Refining Research Citations. through Context Analysis”, Intelligent Informatics, Advances in Intelligent Systems and Computing, Volume 182, 2013, pp 65-71

[17] G.S. Mahalakshmi, S. Sendhilkumar, Optimising Research Progress Trajectories with Semantic Power Graphs, Springer LNCS, in proc. of PREMI 2013, Vol. 8251, pp 708-713.

[18] 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.

[19] 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.

[20] Hirsch JE, "An index to quantify an individual's scientific research output,". Proceedings of the National Academy of Science, vol.102,issue 46, pp.16569-16572, Nov 15 2005.

[21] Iain D. Craig, Andrew M. Plume, Marie E. Mc Veigh, James Pringle, Mayur Amin, Do open access articles have greater citation impact? A Criticla review of the literature, Journal of Informetrics, (1), 2007, 239-248.

[22] Isola Ajiferuke, Kun Lu, Dietmar Wolfram: A comparison of citer and citation-based measure outcomes for multiple disciplines. JASIST 61(10): 2086-2096 (2010)

[23] J. Deepika, G. S. Mahalakshmi, "Journal Impact Factor: A Measure of Quality or Popularity?”IICAI 2011, pp.1138-1157,2011a.

[24] J. M. Kleinberg, “Authoritative sources in a hyperlinked environment”, Journal of ACM, Vol.46, No.5, pp.604–632, 1999.

[25] James H. Flower, Dag W. Aksnes, Does self-citation pay?, Journal of Scientometrics (72), 2007, 427-437.

[26] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank Citation Ranking: Bringing Order to the Web,” technical report,

Stanford Digital Library Technologies Project, 1998.

[27] Michael J. Moravcsik and Poovanalingan Murugesan. 1975. Some results on the function and quality of citations. Social Studies of Science, 5:88--91

[28] S. Sendhilkumar and G.S. Mahalakshmi, Context based Citation Retrieval, International Journal of Networking and Virtual Organizations (IJNVO): Special issue on Open source Intelligence and Web Mining, Inderscience Publishers, Vol. 8, No. 1 / 2, 2011, pp.98-122.

[29] S. Sendhilkumar, Dilip sam and Mahalakshmi G.S. Enhancement of Co-authorship networks with content similarity information, International Conference on Advances in Computing, Communications and Informatics, ICACCI, 2012 , ACM digital library, pp.1225-1228.

[30] S. Sendhilkumar, Mahalakshmi G.S., Harish S., Karthik R., Jagadish M. and Dilip Sam, Assessing Novelty of Research Articles using Fuzzy Cognitive Maps, Intelligent Informatics, Advances in Intelligent Systems and Computing, Springer, Vol. 182, 2013, pp.73-79.

[31] S. Sendhilkumar, Nachiyar S. Nandhini and G.S. Mahalakshmi, Novelty Detection via Topic Modeling in Research Articles, in proceedings of International conference ICCSEA 2013, David C. Wyld (Eds) : ICCSEA, SPPR, CSIA, WimoA, SCAI - 2013 , Vol – 3,pp. 401–410, 2013

[32] S. Sendhilkumar, Elakkiya E. and G.S. Mahalakshmi, Citation Semantic Based Approaches to Identify Article Quality, in proceedings of International conference ICCSEA 2013, David C. Wyld (Eds) : ICCSEA, SPPR, CSIA, WimoA, SCAI - 2013 , Vol – 3,pp. 411–420, 2013

[33] 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.

[34] 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.

[35] Son Bao Pham, Achim Hoffmann , A New Approach for Scientific Citation Classification Using Cue Phrases, AI 2003: Advances in Artificial Intelligence , Lecture Notes in Computer Science Volume 2903, 2003, pp 759-771

[36] Stefano Baccianella, Andrea Esuli and Fabrizio Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining, In Proceedings of the Seventh International Conference on Language Resources and Evaluation, (LREC'10), Valletta, Malta, May 2010

[37] W.B. Lievers, A.K. Pilkey, Characterizing the frequency of repeated citations: The effects of journal, subject area, and self citation, Journal of information Processing and Management (48), 2012, 1116-1123.

[38] Y. Zhang, J. Callan, and T. Minka, “Novelty and redundancy detection in adaptive filtering,” SIGIR '02 Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 81-88, 2002.

ANNEXURE:

Table 1. Articles and Self Citation Articles of Ying Dar Lin

Article ID	Article
A1	I-Wei Chen, Po-Ching Lin, Tsung-Huan Cheng, Chi-Chung Luo, Ying-Dar Lin, Yuan-Cheng Lai, Frank C. Lin: Extracting Ambiguous Sessions from Real Traffic with Intrusion Prevention Systems. I. J. Network Security 14(5): 243-250 (2012)
A2	Ying-Dar Lin, Chia-Yu Ku, Yuan-Cheng Lai, Chia-Fon Hung: In-Kernel Relay for Scalable One-to-Many Streaming. IEEE MultiMedia (IEEEMM) 20(1):69-79 (2013)
A3	Ying-Dar Lin, Po-Ching Lin, Tsung-Huan Cheng, I-Wei Chen, Yuan-Cheng Lai: Low-storage capture and loss recovery selective replay of real flows. IEEE Communications Magazine (CM) 50(4):114-121 (2012)
A4	Ying-Dar Lin, Erica Johnson, Eduardo Joo: Network testing series [Series Editorial]. IEEE Communications Magazine (CM) 50(3):138-139 (2012)
A5	Chun-Nan Lu, Ying-Dar Lin, Chun-Ying Huang, Yuan-Cheng Lai: Session Level Flow Classification by Packet Size Distribution and Session Grouping. AINA Workshops 2012:221-226
A6	Ying-Dar Lin, Chi-Heng Chou, Yuan-Cheng Lai, Tse-Yau Huang, Simon Chung, Jui-Tsun Hung, Frank C. Lin: Test coverage optimization for large code problems. Journal of Systems and Software (JSS) 85(1):16-27 (2012)
A7	Ying-Dar Lin, Erica Johnson, Eduardo Joo: Topics in network testing. IEEE Communications Magazine (CM) 50(9):162 (2012)
A8	Ying-Dar Lin, Chien-Chao Tseng, Cheng-Yuan Ho, Yu-Hsien Wu: How NAT-compatible are VoIP applications? IEEE Communications Magazine (CM) 48(12):58-65 (2010)
A9	Ying-Dar Lin, Shun-Lee Chang, Jui-Hung Yeh, Shau-Yu Cheng: Indoor deployment of IEEE 802.11s mesh networks: Lessons and guidelines. Ad Hoc Networks (ADHOC) 9(8):1404-1413 (2011)
A10	Cheng-Yuan Ho, Fu-Yu Wang, Chien-Chao Tseng, Ying-Dar Lin: NAT-Compatibility Testbed: An Environment to Automatically Verify Direct Connection Rate. IEEE Communications Letters (ICL) 15(1):4-6 (2011)
A11	Cheng-Yuan Ho, Chien-Chao Tseng, Fu-Yu Wang, Jui-Tang Wang, Ying-Dar Lin: To Call or To Be Called Behind NATs is Sensitive in Solving the Direct Connection Problem. IEEE Communications Letters (ICL) 15(1):94-96 (2011)
A12	Yi-Neng Lin, Ying-Dar Lin, Yuan-Cheng Lai, Che-Wen Wu: Highest Urgency First (HUF): A latency and modulation aware bandwidth allocation algorithm for WiMAX base stations. Computer Communications (COMCOM) 32(2):332-342 (2009)
A13	Ying-Dar Lin, Szu-Hao Chen, Po-Ching Lin, Yuan-Cheng Lai: Designing and evaluating interleaving decompressing and virus scanning in a stream-based mail proxy. Journal of Systems and Software (JSS) 81(9):1517-1524 (2008)
A14	Ying-Dar Lin, Ching-Ming Tien, Shih-Chiang Tsao, Ruo-Hua Feng, Yuan-Cheng Lai: Multiple-resource request scheduling for differentiated QoS at website gateway. Computer Communications (COMCOM) 31(10):1993-2004 (2008)
A15	Shih-Chiang Tsao, Yuan-Cheng Lai, Le-Chi Tsao, Ying-Dar Lin: On applying fair queuing discipline to schedule requests at access gateway for downlink differential QoS. Computer Networks (CN) 52(18):3392-3404 (2008)
A16	Yi-Neng Lin, Ying-Dar Lin, Yuan-Cheng Lai: Thread allocation in CMP-based multithreaded network processors. Parallel Computing (PC) 36(2-3):104-116 (2010)
A17	Yi-Neng Lin, Ying-Dar Lin, Yuan-Cheng Lai: Thread allocation in CMP-based multithreaded network processors. Parallel Computing (PC) 36(2-3):104-116 (2010)
A18	Po-Ching Lin, Ying-Dar Lin, Yuan-Cheng Lai: A Hybrid Algorithm of Backward Hashing and Automaton Tracking for Virus Scanning. IEEE Trans. Computers (TC) 60(4):594-601 (2011)
A19	Ying-Dar Lin, Jui-Hung Yeh, Tsung-Hsien Yang, Chia-Yu Ku, Shiao-Li Tsao, Yuan-Cheng Lai: Efficient dynamic frame aggregation in IEEE 802.11s mesh networks. Int. J. Communication Systems (IJCOMSYS) 22(10):1319-1338 (2009)
A20	Yuan-Cheng Lai, Ying-Dar Lin, Fan-Cheng Wu, Tze-Yau William Huang, Frank C. Lin: Embedded TaintTracker: Lightweight Run-Time Tracking of Taint Data against Buffer Overflow Attacks. IEICE Transactions (IEICET) 94-D(11):2129-2138 (2011)

A21	Yi-Neng Lin, Ying-Dar Lin, Yuan-Cheng Lai, Che-Wen Wu: Highest Urgency First (HUF): A latency and modulation aware bandwidth allocation algorithm for WiMAX base stations. Computer Communications (COMCOM) 32(2):332-342 (2009)
A22	Ying-Dar Jason Lin, Horng-Zhu Lai, Yuan-Cheng Lai: A Hierarchical Network Storage Architecture for Video-on-Demand Services. LCN 1996:355-364
A23	Joe Shang-Chieh Wu, Ying-Dar Lin: A novel pairing algorithm for high-speed large-scale switches. IEEE Communications Letters (ICL) 4(1):23-25 (2000)
A24	Ying-Dar Lin, Ren-Kuei Yang, Chi-Chun Lo: Alarm correlation for congestion diagnosis in ATM networks. NOMS 1996:624-627
A25	Ying-Dar Jason Lin, Wei-Ming Yin, Chen-Yu Huang: An Investigation into HFC MAC Protocols: Mechanisms, Implementation, and Research Issues. IEEE Communications Surveys and Tutorials (COMSUR) 3(3):2-13 (2000)
A26	Ying-Dar Jason Lin, Tian-Ren Huang, Yuan-Cheng Lai: Characterization and Control of Highly Correlated Traffic in High-Speed Networks.LCN 1996:19-27
A27	I-Wei Chen, Po-Ching Lin, Chi-Chung Luo, Tsung-Huan Cheng, Ying-Dar Lin, Yuan-Cheng Lai, Frank C. Lin: Extracting Attack Sessions from Real Traffic with Intrusion Prevention Systems. ICC 2009:1-5
A28	Ying-Dar Jason Lin, Tzu-Chieh Tsai, San-Chiao Huang, Mario Gerla: HAP: A New Model for Packet Arrivals. SIGCOMM 1993:212-223
A29	Ying-Dar Lin, Po-Ching Lin, Yuan-Cheng Lai, Tai-Ying Liu: Hardware-Software Codesign for High-Speed Signature-based Virus Scanning.IEEE Micro (MICRO) 29(5):56-65 (2009)
A30	Ying-Dar Jason Lin, Mario Gerla: Induction and Deduction for Autonomous Networks. IEEE Journal on Selected Areas in Communications (JSAC) 11(9):1415-1425 (1993)
A31	Ying-Dar Jason Lin, Yu-Ching Hsu: Multihop Cellular: A New Architecture for Wireless Communications. INFOCOM 2000:1273-1282
A32	Ying-Dar Jason Lin: On IEEE 802.14 Medium Access Control Protocols. IEEE Communications Surveys and Tutorials (COMSUR) 1(1) (1998)
A33	Ying-Dar Jason Lin, Chia-Jen Wu, Wei-Ming Yin: PCUP: Pipelined Cyclic Upstream Protocol over Hybrid Fiber Coax. INFOCOM 1997:1165-1173
A34	Yuan-Cheng Lai, Ying-Dar Lin, Wei-Che Yu, Yuh-Tay Lin: GMNF-DVMRP : A Modified Version of Distance Vector Multicast Routing Protocol. ICCCN 1997:65-69
A35	Yuan-Cheng Lai, Ying-Dar Jason Lin: Performance Analysis of Rate-Based Flow Control under a Variable Number of Sources. Broadband Communications 1999:445-454
A36	Wei-Ming Yin, Ying-Dar Lin: Statistically optimized minislot allocation for initial and collision resolution in hybrid fiber coaxial networks. IEEE Journal on Selected Areas in Communications (JSAC) 18(9):1764-1773 (2000)
A37	Kuo-Kun Tseng, Ying-Dar Lin, Tsern-Huei Lee, Yuan-Cheng Lai: A Parallel Automaton String Matching with Pre-Hashing and Root-Indexing Techniques for Content Filtering Coprocessor. ASAP 2005:113-118
A38	Ying-Dar Lin, Kuo-Kun Tseng, Tsern-Huei Lee, Yi-Neng Lin, Chen-Chou Hung, Yuan-Cheng Lai: A platform-based SoC design and implementation of scalable automaton matching for deep packet inspection. Journal of Systems Architecture (JSA) 53(12):937-950 (2007)
A39	Po-Ching Lin, Ming-Dao Liu, Ying-Dar Lin, Yuan-Cheng Lai: An Early Decision Algorithm to Accelerate Web Content Filtering. ICOIN 2006:833-841
A40	Huan-Yun Wei, Shih-Chiang Tsao, Ying-Dar Jason Lin: Assessing and Improving TCP Rate Shaping over Edge Gateways. IEEE Trans. Computers (TC) 53(3):259-275 (2004)
A41	Huan-Yun Wei, Ching-Chuang Chiang, Ying-Dar Lin: Co-DRR: An Integrated Uplink and Downlink Scheduler for Bandwidth Management over Wireless LANs. IEICE Transactions (IEICET) 90-B(8):2022-2033 (2007)
A42	Ying-Dar Lin, Chih-Wei Jan, Po-Ching Lin, Yuan-Cheng Lai: Designing an Integrated Architecture for Network Content Security Gateways.IEEE Computer (COMPUTER) 39(11):66-72 (2006)

A43	Kuo-Kun Tseng, Ying-Dar Lin, Tsern-Huei Lee, Yuan-Cheng Lai: Deterministic high-speed root-hashing automaton matching coprocessor for embedded network processor. SIGARCH Computer Architecture News (SIGARCH) 35(3):36-43 (2007)
A44	Ying-Dar Lin, Po-Ching Lin, Meng-Fu Tsai, Tsao-Jiang Chang, Yuan-Cheng Lai: kP2PADM: An In-kernel Gateway Architecture for Managing P2P Traffic. IPDPS 2007:1-9
A45	Po-Ching Lin, Zhi-Xiang Li, Ying-Dar Lin, Yuan-Cheng Lai, Frank C. Lin: Profiling and accelerating string matching algorithms in three network content security applications. IEEE Communications Surveys and Tutorials (COMSUR) 8(1-4):24-37 (2006)
A46	Ying-Dar Lin, Ching-Ming Tien, Shih-Chiang Tsao, Ruo-Hua Feng, Yuan-Cheng Lai: Multiple-Resource Request Scheduling for Differentiated QoS at Website Gateway. AINA 2008:433-440
A47	Yi-Neng Lin, Yao-Chung Chang, Ying-Dar Lin, Yuan-Cheng Lai: Resource allocation in network processors for network intrusion prevention systems. Journal of Systems and Software (JSS) 80(7):1030-1036 (2007)
A48	Ying-Dar Lin, Kuo-Kun Tseng, Chen-Chou Hung, Yuan-Cheng Lai: Scalable Automaton Matching for High-Speed Deep Content Inspection. AINA Workshops 2007:858-863
A49	Ching-Ming Tien, Cho-Jun Lee, Po-Wen Cheng, Ying-Dar Lin: SOAP Request Scheduling for Differentiated Quality of Service. WISE Workshops 2005:63-72
A50	Shih-Chiang Tsao, Yuan-Cheng Lai, Ying-Dar Lin: Taxonomy and Evaluation of TCP-Friendly Congestion-Control Schemes on Fairness, Aggressiveness, and Responsiveness. IEEE Network (NETWORK) 21(6):6-15 (2007)
A51	Yi-Neng Lin, Chiuan-Hung Lin, Ying-Dar Lin, Yuan-Cheng Lai: VPN Gateways over Network Processors: Implementation and Evaluation.IEEE Real-Time and Embedded Technology and Applications Symposium 2005:480-486
A52	Huan-Yun Wei, Ying-Dar Jason Lin: A Survey and Measurement-Based Comparison of Bandwidth Management Techniques. IEEE Communications Surveys and Tutorials (COMSUR) 5(2):10-21 (2003)
A53	Po-Ching Lin, Ming-Dao Liu, Ying-Dar Lin, Yuan-Cheng Lai: Accelerating Web Content Filtering by the Early Decision Algorithm. IEICE Transactions (IEICET) 91-D(2):251-257 (2008)
A54	Yu-Ching Hsu, Ying-Dar Lin: Base-centric routing protocol for multihop cellular networks. GLOBECOM 2002:158-162
A55	Tsung-Huan Cheng, Ying-Dar Lin, Yuan-Cheng Lai, Po-Ching Lin: Evasion Techniques: Sneaking through Your Intrusion Detection/Prevention Systems. IEEE Communications Surveys and Tutorials (COMSUR) 14(4):1011-1020 (2012)
A56	Yuan-Cheng Lai, Ying-Dar Jason Lin, Chih-Yu Chen, Huan-Yun Wey: Guaranteed versus Controlled Load: Implications for Service Subscribers and Providers in RSVP Networks. ICOIN 2001:487-494
A57	Ying-Dar Lin, Po-Ching Lin, Meng-Fu Tsai, Tsao-Jiang Chang, Yuan-Cheng Lai: kP2PADM: An In-kernel Gateway Architecture for Managing P2P Traffic. IPDPS 2007:1-9
A58	Yi-Neng Lin, Ying-Dar Lin, Kuo-Kun Tseng, Yuan-Cheng Lai: Modeling and analysis of core-centric network processors. ACM Trans. Embedded Comput. Syst. (TECS) 8(2) (2009)
A59	Ying-Dar Lin, Ching-Ming Tien, Shih-Chiang Tsao, Ruo-Hua Feng, Yuan-Cheng Lai: Erratum to "Multiple-resource request scheduling for differentiated QoS at website gateway". Computer Communications (COMCOM) 31(17):4230 (2008)
A60	Kuo-Kun Tseng, Yuan-Cheng Lai, Ying-Dar Lin, Tsern-Huei Lee: A fast scalable automaton-matching accelerator for embedded content processors. ACM Trans. Embedded Comput. Syst. (TECS) 8(3) (2009)
A61	Ying-Dar Lin, Shiao-Li Tsao, Shun-Lee Chang, Shau-Yu Cheng, Chia-Yu Ku: Design issues and experimental studies of wireless LAN Mesh. IEEE Wireless Commun. (WC) 17(2):32-40 (2010)
A62	Tsung-Huan Cheng, Ying-Dar Lin, Yuan-Cheng Lai, Po-Ching Lin: Evasion Techniques: Sneaking through Your Intrusion Detection/Prevention Systems. IEEE Communications Surveys and Tutorials (COMSUR) 14(4):1011-1020 (2012)
A63	I-Wei Chen, Po-Ching Lin, Chi-Chung Luo, Tsung-Huan Cheng, Ying-Dar Lin, Yuan-Cheng Lai, Frank C. Lin: Extracting Attack Sessions from Real Traffic with Intrusion Prevention Systems. ICC 2009:1-5

A64	Ying-Dar Lin, Po-Ching Lin, Yuan-Cheng Lai, Tai-Ying Liu: Hardware-Software Codesign for High-Speed Signature-based Virus Scanning.IEEE Micro (MICRO) 29(5):56-65 (2009)
A65	Ying-Dar Lin, I-Wei Chen, Po-Ching Lin, Chang-Sheng Chen, Chun-Hung Hsu: On campus beta site: architecture designs, operational experience, and top product defects. IEEE Communications Magazine (CM) 48(12):83-91 (2010)
A66	Chia-Yu Ku, Ying-Dar Lin, Yuan-Cheng Lai, Pei-Hsuan Li, Kate Ching-Ju Lin: Real traffic replay over WLAN with environment emulation.WCNC 2012:2406-2411
A67	Po-Ching Lin, Ying-Dar Lin, Yuan-Cheng Lai, Yi-Jun Zheng, Tsern-Huei Lee: Realizing a Sub-Linear Time String-Matching Algorithm With a Hardware Accelerator Using Bloom Filters. IEEE Trans. VLSI Syst. (TVLSI) 17(8):1008-1020 (2009)
A68	Yi-Neng Lin, Ying-Dar Lin, Yuan-Cheng Lai: Thread Allocation in Chip Multiprocessor Based Multithreaded Network Processors. AINA 2008:718-725
A69	Chun-Nan Lu, Chun-Ying Huang, Ying-Dar Lin, Yuan-Cheng Lai: Session level flow classification by packet size distribution and session grouping. Computer Networks (CN) 56(1):260-272 (2012)
A70	Ying Dar Lin,Chen Yu Huang,Wui-Ming Yin:Allocation and Schleduling Algorithms for 1EE:E 802.14 and MCNS in Hybrid Fiber Coaxial Networks, IEEE Transactions on Broadcasting 44(4):427-435 (1998)
A71	Ying-Dar Lin, Yu-Ching Hsu, Kuan-Wen Oyang, Dong-Su Yang, Tzu-Chieh Tsai, "Multihop Wireless IEEE 802.11 LANs: A Prototype Implementation," Journal of Communications and Networks, 2(4), Dec. (2000).
A72	Ming-Wei Wu, Ying-Dar Jason Lin: Open Source Software Development: An Overview. IEEE Computer (COMPUTER) 34(6):33-38 (2001)
A73	Wei-Ming Yin, Chia-Jen Wu, Ying-Dar Lin, "Two-phase Minislot Scheduling Algorithm for HFC QoS Services Provisioning," IEICE Transactions on 12 Communications,E85-B (3) March(2002).
A74	Ying-Dar Lin, Yi-Neng Lin, Shun-Chin Yang, Yu-Sheng Lin: DiffServ over Network Processors: Implementation and Evaluation. Hot Interconnects 2002:121-126
A75	Lin, Y. and W.S. Wong, V." Frame Aggregation and Optimal Frame Size Adaptation for IEEE 802.11n WLANs." in Proc. of IEEE Global Telecom. Conf. (2006).
A76	Ying-Dar Jason Lin, Huan-Yun Wei, Shao-Tang Yu: Building an Integrated Security Gateway: Mechanisms, Performance Evaluations, Implementations, and Research Issues. IEEE Communications Surveys and Tutorials (COMSUR) 4(1):2-15 (2002)
A77	Joe Shang-Chieh Wu, Ying-Dar Lin: An efficient and orderly implementation of bypass queue under bursty traffic. Parallel Computing (PC) 24(14):2143-2148 (1998)
A78	Wei Ming Yin, Chia Jen Wu, Ying Dar Lin:Two-Phase minislot Scheduling Algorithm for HFC QoS Services Provisioning, In Global Telecommunications Conference,1,410-414,(2001)
A79	Mario Gerla, Ying-Dar Lin: Network management using database discovery tools. LCN 1991:378-385
A80	Yi-Neng Lin,Shih-Hsin Chien,Ying-Dar Lin,Yuan-Cheng Lai,Mingshou Liu :DYNAMIC BANDWIDTH ALLOCATION FOR 802.16E-2005 MAC, In Current Technology Developments of WiMax Systems pp 17-29 ,(2009)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

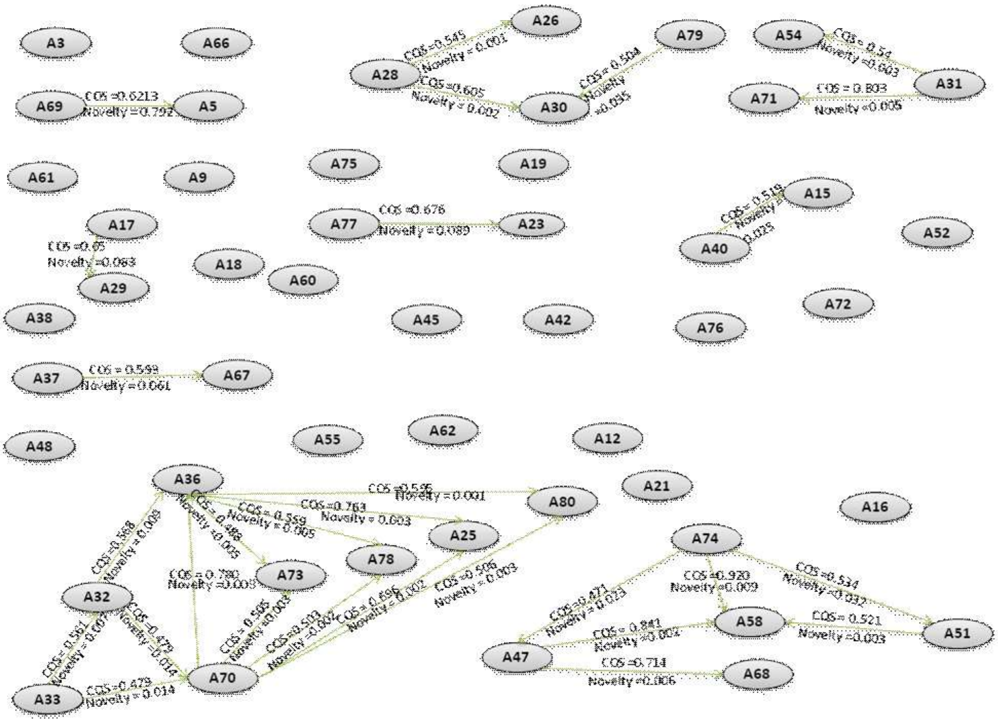

Figure 4: Reduced Author Self Citation Network based on SCQS