Lessons Learned: The Complexity of Accurate Identification of in-Text Citations

Abdul Shahid, Muhammad Tanvir Afzal, and Muhammad Abdul Qadir
Department of Computer Science, Mohammad Ali Jinnah University, Pakistan

Abstract: *The importance of citations is widely recognized by the scientific community. Citations are being used in making a number of vital decisions such as calculating impact factor of journals, calculating impact of a researcher (H-Index), ranking universities and research organizations. Furthermore, citation indexes, along with other criteria, employ citation counts to retrieve and rank relevant research papers. However, citing patterns and in-text citation frequency are not used for such important decisions. The identification of in-text citation from a scientific document is an important problem. However, identification of in-text citation is a tough ask due to the ambiguity between citation tag and content. This research focuses on in-text citation analysis and makes the following specific contributions such as: Provides detailed in-text citation analysis on 16,000 citations of an online journal, reports different pattern of citations-tags and its in-text citations and highlights the problems (mathematical ambiguities, wrong allotments, commonality in content and string variation) in identifying in-text citations from scientific documents. The accurate identification of in-text citations will help information retrieval systems, digital libraries and citation indexes.*

Keywords: *In-text citation analysis, citation frequency, citation tag, in-text citation patterns, digital library.*

Received June 27, 2013; accepted March 19, 2014; published online December 3, 2014

1. Introduction

Citations play an important role in making a number of important decisions such as calculating impact factor of journals [6], calculating impact of a researcher (H-Index) [10], ranking universities [9] and research organizations. Furthermore, citation indexes, along with other criteria, employ citation counts to retrieve and rank relevant research papers [3, 7]. Currently, the existing tools make use of citation counts from "Reference" section. Citations count refer to the number of times a research document has been cited by other research documents, while in-text citation frequency means how many times a cited article is being referred in the text of the cited-by article.

The importance of in-text citation frequency has been realized by number of researchers. For example, Teufel [21] concluded that the performance of semantic analysis systems can be improved by analyzing citation tag in different sections of an article [20, 21]. Maricic *et al.* [15] analyzed in-text citation behaviours for evaluating the current practices of scientific achievements. Gipp *et al.* [8, 14] have proposed to extend state-of-the-art system (co-citation) by exploiting in-text citations. They analyzed the distribution of co-citation behaviours in cited by documents for discovering related documents. Shahid *et al.* [18] used in-text citation patterns to discover implicit relationships between cited and cited-by documents. However, none of these approaches have discussed how in-text citation can be discovered.

There is an online tool named as PDFX that is specifically designed for the conversion of scientific documents from PDF format to XML format. We have manually inspected its results and found that the identification of in-text citation is a tough ask due to reasons such as: Conversions from PDF to text and ambiguity between citation tags and contents of a document.

In this paper, we report pragmatic in-text citation analysis on the insights of in-text citation occurrences in the text of a scientific document. There are a number of scientific writing templates for referencing such as IEEE, Harvard, Vancouver and APA etc. Each referencing template has its own way to cite other's work. We discovered that there are certain scenarios in each template where the identification of in-text citation is not accurate. Furthermore, this research reports on a number of problems identified which are associated with ambiguity between citation tag and content of document belonging to different referencing templates. There are also other problems which are due to the conversion from PDF to text (special encoding documents and figures etc., are not properly converted) [2] and thus, make the in-text citation identification a daunting ask. During the experiment, we confronted certain problems which are harder and need careful attention while extracting in-text citations. These problems are discussed in section 4.

There are mainly two approaches used for conversion from PDF to text. One approach is extracting text from PDF document and keeping .txt version of the document. The other approach extracts text from PDF document and creates its XML (structured) version. We evaluated number of text extraction tools. Based on our manual inspection of the extracted results, we selected a tool named as PDFX[1]. The PDFX [5] is specifically built for the conversion of scientific documents from PDF to XML. The rest of the paper is organized as follows: Section 2 describes the related work. Section 3 presents research methodology and in-text citation analysis. Section 4 demonstrates various problems that need careful consideration, followed by the summary and conclusion sections.

2. Related Work

In-text citations have been used in literature for various purposes. For example, Teufel [21] discussed that exploiting in-text citations in sections of scientific documents could provide better results for determining the sentiments of the citing author for a cited article. For example when citation occurs in "Introduction" section, it will probably be for a supporting document. In a similar way, the citations in the "Results" section will have significance rather than the citations in the "Related Work" section. In this direction, recently, Ciancarini *et al.* [4, 11, 12, 13] have proposed a technique namely CiTalO, a tool for inferring citation function by performing various steps such as: Ontology learning, sentiment analysis, word-sense disambiguation and ontology mapping. Their proposed approach exploits text around in-text citation occurrence to discover citation relationship. The proposed approach automatically annotates citations with properties defined in Citation Typing Ontology (CiTO)[2]. They have tested their proposed system on a very small dataset of just 18 papers, of them, total 377 citations were retrieved. They have evaluated their system using precision and recall commonly used in information retrieval for evaluation purposes [16]. The precision and recall of their system is not good i.e., below 50 percent due to various reasons such as: Coverage of CiTO properties, noise of proximity synsets, matching synsets and compound word properties and identification of the context window of citations etc.

Maricic *et al.* [15] conducted an analysis of citations by exploiting the location of citations. They performed experiments on a large corpus produced by a multidisciplinary institute. They recorded the level of citing as high or low according to the frequency of in-text citations. They classified a cited document as the essential, central or organic citing type based on meaningful or high intensity of citing level. Similarly, they classified a cited paper as the non-essential, peripheral, or perfunctory citing type based on cursory or low intensity citing level. The results show that the cursory or low intensity citations were dominant in the introduction section of papers and meaningful or high intensity citations were dominant in methods, results, and discussion sections. Finally, the results were analyzed and they discovered that the results do not support the current practice of evaluating scientific achievements by means of statistical analysis of citation counts without considering additional information.

Gipp *et al.* [8, 14] have analyzed distributions of co-citations at four levels of proximity (such as: Journal articles, organizational sections in articles, co-citation frequency groups and roles of co-citations) with reference to corresponding traditional co-citation network. They found that sentence level co-citations play a predominant role in forming the overall co-citation network. Their results indicated that sentence level co-citations are more useful candidates for co-citation analysis because they tend to preserve the essential structural components of the corresponding traditional co-citation network. Furthermore, sentence level co-citations tend to appear much infrequent in comparison to loosely coupled article level co-citations.

Shahid *et al.* [18] reported that in-text citations can be used as a measure to determine relationship between cited and cited-by article. They empirically proved that if cited-by paper cites a paper more than five times (in-text frequency=5), then there exist a strong relationship between cited by and cited paper.

However, none of the above research contributed towards the problems and difficulties of accurate identification of citation tag in the text of the cited-by document. At first glance, it looks a trivial task to identify in-text citation frequencies, reality is different. We discuss few real scenarios which shed light on the difficulty and challenging nature of accurate identification of citation tag.

3. Research Methodology

To perform detailed in-text citation analysis, we used XML version of the documents. PDF document were converted into xml using PDFX. Subsequently, we formulated queries by using XPath/XQuery expressions to compute in-text citation frequencies of each cited article in cited by article. In-text citation frequencies were grouped based on citation frequencies so that detailed analysis of correct and incorrect in-text citation identification can be performed. Interesting findings were discovered: Handsome amount of citations were found that were not even referred single time in text of cited by article. Based upon this, we find various reasons of incorrect identification of in-text citation frequencies as explained in following sections.

3.1. PDF to XML Conversion

Current digital libraries, online journals and conferences generally publish and maintain research in PDF format. To perform experimentation over scholarly documents, we need to extract text from PDF documents. There are mainly two approaches used for

[1] http://pdfx.cs.man.ac.uk/

[2] http://www.essepuntato.it/lode/http://purl.org/spar/cito

storing extracted text from PDF document. One deals with the conversion of PDF to text [1, 2]. The other approach deals with the conversion from PDF to XML [17]. In the first approach (PDF to text), .txt version of the document is created. The most renowned tool in this category is pdfbox[3] tool. In the second approach XML version of the PDF document is created. In this category, there are many commercial tools available that convert a PDF file to XML file e.g., OmniPage[4] and OpenXerox[5]. These tools extract text from PDF file and create XML tags either at character or word level. There is another tool PDFx that is specifically designed for scientific document conversion from PDF to XML. The PDFx extracts the content using SPAR: DOCo [19] ontology and creates structured document XML for a given PDF document. Based on the performance and features, we selected PDFx to convert all available research papers from Journal of Universal Computer Science (J.UCS[6]). The J.UCS was selected for this experiment due to the following reasons: An online and open access journal, it covers all topics of computer science and authors come from diversified research experiences. Therefore, this will help our system to make a comprehensive analysis of in-text citation patterns. There were more than 1,200 PDF documents which were converted from PDF to XML by using PDFX.

3.2. Extracting in-Text Citation Frequencies

When dataset was ready, xPath[7] and xQuery[8] based solution was built to extract in-text citation frequencies. The in-text citation frequency means that how many times a citation has been referred in the text of cited by article. The PDFx tool separately marks each reference and includes citation marker around its in-text citation in text of the article. For example, consider the following example of a typical reference string extracted using PDFx.

"For a comprehensive study of Workflow products and their characteristics see [<xref ref-type="bibr" rid="R2" id="84" class="deo:Reference">Aalst, Hee 02</xref>]".

In this text snippet the "rid="R2"" means that it is reference number 2 in the reference list and its citation tag was Identified as "Aalst, Hee 02". The xPath and xQuery expressions were designed to extract each reference and its in-text citation frequencies. Furthermore, the results were persisted in relational database for further analysis. Approximately 16,000 citations along with in-text citations were retrieved automatically and were used for further analysis.

3.3. Grouping Papers Based on in-Text Citation Frequency

From the persisted data, we retrieved in-text citation frequencies between the range of 0 and 22. This means that in the data of 16,000 references, the minimum value of in-text citation frequency was zero and the maximum value of in-text citation frequency was 22.

For the analysis of in-text citation frequency, we segregated the papers into different groups depending upon the value (number of occurrences) of in-text citations. Then, we analyzed the in-text citation frequencies and their patterns belonging to different groups. We made the following six groups: The Group 1 represents all such papers whose in-text citation frequencies were greater than 15, Group 2 denotes all such papers whose in-text citation frequencies were in the range of 10-15, Group 3 stands for all such papers whose in-text citation frequencies were in the range of 7-9, Group 4 represents all such papers whose in-text citation frequencies were in the range of 4-6, Group 5 corresponds to all such papers whose in-text citation frequencies were in the range of 1-3 and the papers whose in-text citation frequencies were zero belong to the Group 6. We manually verified in-text citation frequencies in the body of the cited by papers. For the first three groups, we exhaustively analyzed all in-text citations, however, from the remaining groups, a reasonable sample data was selected for this study. We calculated correct and incorrect in-text citation frequencies manually from the real documents.

The overall results are shown in the Table 1. The First column represents all groups and the correct and and incorrect columns lists the overall percentage of correct and incorrect in-text citation frequencies for each group. The overall accuracy for in-text citation identification is 58%. It means that 42% in-text citations were not properly identified. The Table 1 represents a pattern i.e., with the increase of in-text citation frequency, the incorrect results increases. However, this does not apply to in-text citation frequency equal to zero.

Table 1. Percentage of correct/incorrect marking of in-text citations.

In-Text Citation Frequencies	Correct	Incorrect
>15 (TR: 16)	40	60
10-15 (TR: 50)	44	56
7-9 (TR: 104)	67	33
4-6 (TR: 555)	70	30
1-3 (TR: 12712)	90	10
=0 (TR:5091)	36	64

3.4. References with Zero in-Text Citation Frequency

There is another interesting finding that out of 16,000 citations, we were able to identify more than 3,000 citations which were not even cited a single time in the body text of the cited by document. Such citations are being used for making vital decision such as: Calculating impact factors of journals, H-Index of authors etc., such authoritative systems may cross check the in-text citation frequencies before making

[3] http://pdfbox.apache.org/

[4] http://www.nuance.com/for-business/by product/omnipage/index.htm

[5] http://open.xerox.com/

[6] http://www.jucs.org/

[7] http://www.w3.org/standards/techs/xpath

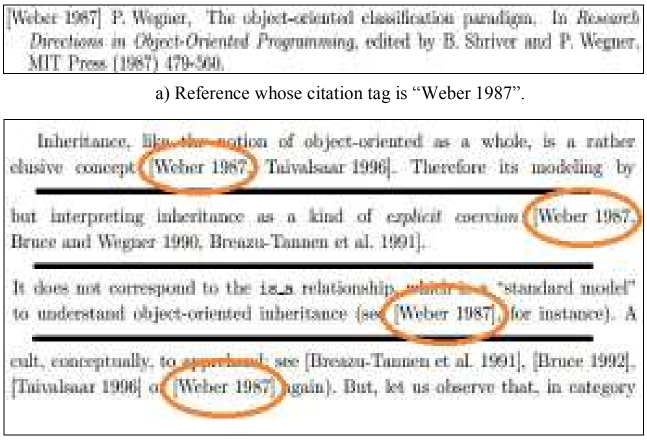
[8] http://www.w3.org/TR/xquery/

such vital decisions. Furthermore, the administration of journals/conference and reviewers should at least make sure that all reference have been cited even once in the body text of the document.

3.5. Identification of in-Text Citation Frequency

A citation tag is a unique combination of characters used to cite a particular reference in the body text of the paper. For example, consider the following scenario: Figure 1-a represents a typical reference from a real document from our dataset where the citation tag is "[Weber 1987]". Figure 1-b represents text snippets where the citation tag "[Weber 1987]" has been used. It is obvious from Figure 1-b that the in-text citation frequency of this reference is four.

In this section we will explain the reasons for incorrect identification of in-text citations. For this purpose, we have identified clusters based on different types of citation tags.



[Weber 1987] P. Wegner, The object-oriented classification paradigm. In *Research Directions in Object-Oriented Programming*, edited by B. Shriver and P. Wegner, MIT Press (1987) 479-560.

a) Reference whose citation tag is "Weber 1987".

Inheritance, like the notion of object-oriented as a whole, is a rather elusive concept [Weber 1987, Taivalsaar 1996]. Therefore its modeling by

but interpreting inheritance as a kind of *explicit* coercion [Weber 1987, Bruce and Wegner 1990, Breazu-Tannen et al. 1991].

It does not correspond to the is_a relationship, which is a "standard model" to understand object-oriented inheritance (see [Weber 1987], for instance). A

cult, conceptually, to apprehend: see [Breazu-Tannen et al. 1991], [Bruce 1992], [Taivalsaar 1996] or [Weber 1987] again). But, let us observe that, in category

b) In-text citations of "Weber 1987" in body of the article.

Figure 1. Typical scenario of in-text citation occurrences for a reference

3.5.1. Clustering Citation-Tags

The clusters based on citation-tags are as follows:

- Numeric: This cluster represents all such citations which have a numeric citation tag for example "[1]", "1." and "(1)" etc.
- Alphabetic: This cluster represents all such citations which have an alphabetic citation tag. This cluster was the most populated one. The citation-tag examples of this cluster are: "Srinivasan, Scherbakov 1995", [Davenport and Prusak, 1998], [Staiger 1993], [Olson *et al.* 2002], [MPEG-7] etc.
- Single Character: This is an interesting cluster having citation-tags of single character long such as "[N]", "[P]", "[A]" etc. However, this cluster was the less populated one.

3.5.2. Identification of Incorrect in-Text Citations

We identified a number of different reasons for incorrect identification of in-text citations as listed all below. Each reason has a relation with the above mentioned clusters:

- Wrong Allotment: When in-text citation of one cited article is assigned to another cited article.
- Mathematical Ambiguities: When Intervals, equation, figures or vector values are considered as in-text citations.
- Commonality in Content: When normal text is considered as in-text citation tag. For example, we have a citation tag [P] of a reference and "P" is very common character which is being used in the paper's content frequently.
- String Variations: When the citation tag in the text of the document is a variant of citation tag in the reference list. These variations are normally due to include/exclude of some characters. Sometimes, authors may refer a citation bit differently in the content as compared to the reference. For example, the citation tag from the reference list "[Davenport and Prusak, 1998]" may be referred in the text of the document in different ways such as: "[Daven-port and Prusak, 1998]", "[Davenport and Prusak, 1998]" and "[Davenport-and Prusak, 1998]" etc.

The overall results are presented in Figure 2. The X-axis shows different clusters as discussed above. The Y-axis shows error percentage in different categories. This graph shows interesting patterns, for example, the error category "commonality in Content" is the most frequently occurring category in the cluster "Single Character"., the "String Variations" and "Wrong Allotment" are related with the cluster "Alphabetic", and the "Mathematical ambiguities" is the most highlighted problem in the cluster "Numeric", however, "String variations" is also an important issue to be addressed in the cluster "Numeric".

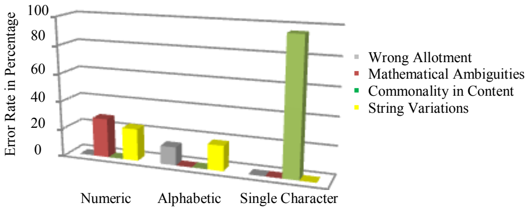


Figure 2. Reasons for wrong identification of in-text citations.

This comprehensive study of more than 16,000 citations identified insights in the identification of in-text citations. This analysis is helpful for the systems which identifies in-text citations. The error categories are strongly correlated with the clusters. For example, if a citation entry has an Alphabetic citation tag, the system should focus on the issues of "Wrong allotment" and "String variations".

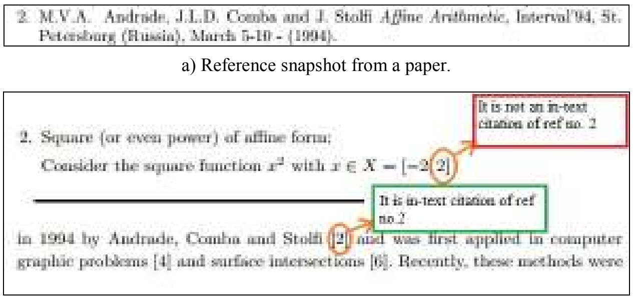
4. Real Scenarios From Scientific Documents

Based on manual inspection and analysis of the incorrect results, we are presenting interesting real scenarios from the documents where in-text citation has been identified incorrectly.

The following scenarios demonstrate real issues where accurate identification of in-text citations is problematic. These scenarios highlight the ambiguity of identification of citation tags in a typical part of paper's content. Below is the detail of each scenario. Each scenario is a typical example of common reasons identified above.

4.1. Scenario 1-Mathematical Ambiguity Interval

A reference is shown in Figure 3-a extracted from reference sections of an article. In this case, the citation tag is "2". The citation in the running text of the document could be made using the following citation tags: "[2]", "[2,", ", 2]", "[2", "2]". "[, 2,]" or it can be hidden in the following citation tag "[1-5]" which is referring all references from 1 to 5. However, Figure 3-b presents another snippet from the same document where "[-2, 2]" is part of the paper text and does not belong to a citation tag. The tag "[-2, 2]" is being used in a mathematical formula for donating an interval. Traditional in-text citation discovery systems will incorrectly make this interval values as in-text citation of reference "2".



2. M.V.A. Andrade, J.L.D. Comba and J. Stolfi *Affine Arithmetic*, Interval'94, St. Petersburg (Russia), March 5-10 - (1994).

a) Reference snapshot from a paper.

It is not an in-text citation of ref no. 2

2. Square (or even power) of affine form;

Consider the square function x^2 with $x \in X = [-2, 2]$

It is in-text citation of ref no.2

in 1994 by Andrade, Comba and Stolfi [2] and was first applied in computer graphic problems [4] and surface intersections [6]. Recently, these methods were

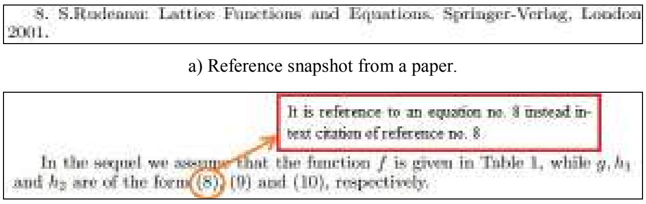
b) Content snippet that can mislead the results for above reference.

Figure 3. Scenario-1: mathematical ambiguity interval.

For tackling this type of problems, the automated tool needs to discover the context of the citation and needs to disambiguate between actual citation tag and content of the paper.

4.2. Scenario 2-Mathematical Ambiguity Parenthesis

This scenario is an extension of the scenario number 1. A reference is shown in Figure 4-a from the reference section of an article where its citation tag is "8". In the body text of that article, "(8)" could be the one possible citation tag. However, Figure 4-b demonstrates a text from the same document where the "(8)" is being referred for some mathematical equation defined in that article. Thus, it will again become ambiguous for an automated tool to identify in-text citation accurately. Similarly, another example of mathematical ambiguity is shown in Figure 5-a and Figure 5-b. In the shown example the citation tag "[1]" is used to refer first reference. However, in body text of the paper there are some assertion being made and referred as (1). Therefore, again it will become ambiguous for automated tools to correctly mark in-text citation for that reference "[1]".



8. S.Rudeanu: Lattice Functions and Equations. Springer-Verlag, London 2001.

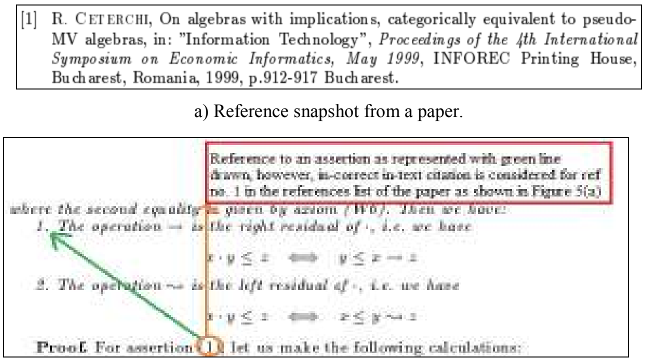
a) Reference snapshot from a paper.

It is reference to an equation no. 8 instead in-text citation of reference no. 8

In the sequel we assume that the function f is given in Table 1, while g, h_1 and h_2 are of the form (8), (9) and (10), respectively.

b) Content snippet that can mislead the results.

Figure 4. Scenario-2: mathematical ambiguity interval.



[1] R. CETERCHI, On algebras with implications, categorically equivalent to pseudo-MV algebras, in: "Information Technology", *Proceedings of the 4th International Symposium on Economic Informatics, May 1999*, INFOREC Printing House, Bucharest, Romania, 1999, p.912-917 Bucharest.

a) Reference snapshot from a paper.

Reference to an assertion as represented with green line drawn, however, in-correct in-text citation is considered for ref no. 1 in the references list of the paper as shown in Figure 5(a)

where the second equality is given by axiom (W6). Then we have:

1. The operation → is the right residual of ·, i.e. we have

$$x \cdot y \leq z \iff y \leq x \to z$$

2. The operation ⇝ is the left residual of ·, i.e. we have

$$x \cdot y \leq z \iff x \leq y \rightsquigarrow z$$

Proof. For assertion (1) let us make the following calculations:

b) Content snippet that can mislead the results for above reference.

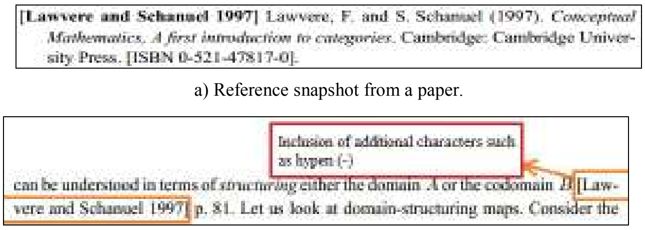
Figure 5. Scenario-2: mathematical ambiguity parenthesis.

The equation number and intervals were found as two important misleading contents for the accurate identification of in-text citation frequencies. These types of problems increased the incorrect results as were shown in the Table 1. These kinds of problem may be addressed by disambiguating in-text citation and context of usage of such citation tag in article.

4.3. Scenario 3-String Variations

In this scenario, we have shown that hyphen can be used within the citation tag while referring to a particular reference in body text of the document. For example, in Figure 6-a, the citation tag is "[Lawvere and Schanuel 1997]", however, Figure 6-b represents a snippet from the same document where the citation tag [Law-vere and Schanuel 1997] is used to refer to that reference. The inclusion of additional characters such as hypen (-) in the in-text citation was another reason.

These types of problems can be resolved using some string comparisons such as edit distance and Levenshtein distance etc.



[Lawvere and Schanuel 1997] Lawvere, F. and S. Schanuel (1997). *Conceptual Mathematics. A first introduction to categories.* Cambridge: Cambridge University Press. [ISBN 0-521-47817-0].

a) Reference snapshot from a paper.

Inclusion of additional characters such as hypen (-)

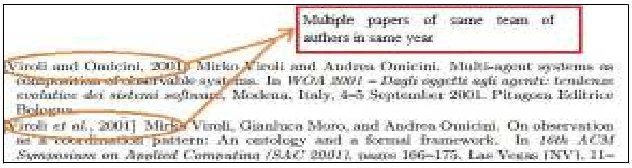
can be understood in terms of *structuring* either the domain A or the codomain B [Law-vere and Schanuel 1997] p. 81. Let us look at domain-structuring maps. Consider the

b) Content snippet that can mislead the results for above reference.

Figure 6. Scenario-3: String variation.

4.4. Scenario 4-Wrong Allotment

In J.UCS dataset we found that some articles have used authors and year information for citation tag. Multiple papers of an author with different team in the same year are referred as shown in Figure 7-a.

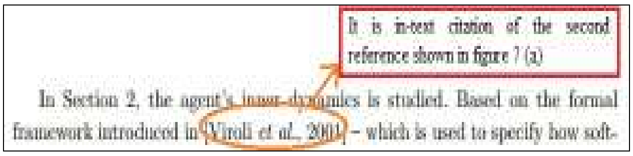


Multiple papers of same team of authors in same year

[Viroli and Omicini, 2001] Mirko Viroli and Andrea Omicini. Multi-agent systems as composition of observable systems. In *WOA 2001 – Dagli oggetti agli agenti: tendenze evolutive dei sistemi software*, Modena, Italy, 4–5 September 2001. Pitagora Editrice Bologna.

[Viroli *et al.*, 2001] Mirko Viroli, Gianluca Moro, and Andrea Omicini. On observation as a coordination pattern: An ontology and a formal framework. In *16th ACM Symposium on Applied Computing (SAC 2001)*, pages 166–175. Las Vegas (NV), 11–

a) Reference snapshot from a paper.



It is in-text citation of the second reference shown in figure 7 (a)

In Section 2, the agent's inner-dynamics is studied. Based on the formal framework introduced in [Viroli et al., 2001] – which is used to specify how soft-

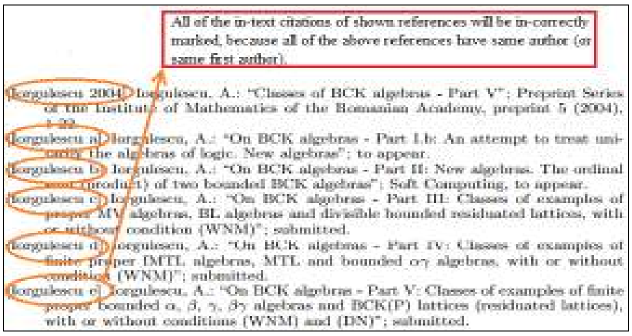
b) Content snippet that can mislead the results.

Figure 7. Scenario-3: Wrong allotment.

There are two separate tags for each citation i.e., [Viroli and Omicine, 2001] and [Viroli *et al.* 2001]. Automated solutions such as PDFx wrongly build a regular expression for citation tag based on only first author and year information.

Therefore, a regular expression, designed to calculate in-text citation of "Viroli, 2001" would mislead the results. Improper building of regular expression was one of the reasons that took part in the overall improper marking of in-text citation as shown in Table 1. To solve such problems, we should design a regular expression carefully such as in the above case, two separate regular expression should be designed: [Viroli and Omicine, 2001] and [Viroli *et al*., 2001].

Similar to above example, in Figure 8, references snapshot from a paper is shown. In this case, automated tools may fails due regular expression for finding in-text citations based on first author of a paper.



All of the in-text citations of shown references will be in-correctly marked, because all of the above references have same author (or same first author).

[Iorgulescu 2004] Iorgulescu, A.: "Classes of BCK algebras - Part V"; Preprint Series of the Institute of Mathematics of the Romanian Academy, preprint 5 (2004), 1-22.

[Iorgulescu a] Iorgulescu, A.: "On BCK algebras - Part I.b: An attempt to treat unitarily the algebras of logic. New algebras"; to appear.

[Iorgulescu b] Iorgulescu, A.: "On BCK algebras - Part II: New algebras. The ordinal sum (product) of two bounded BCK algebras"; Soft Computing, to appear.

[Iorgulescu c] Iorgulescu, A.: "On BCK algebras - Part III: Classes of examples of proper MV algebras, BL algebras and divisible bounded residuated lattices, with or without condition (WNM)"; submitted.

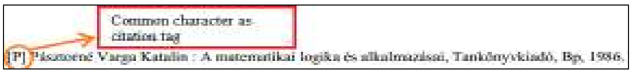
[Iorgulescu d] Iorgulescu, A.: "On BCK algebras - Part IV: Classes of examples of finite proper IMTL algebras, MTL and bounded $\alpha\gamma$ algebras, with or without condition (WNM)"; submitted.

[Iorgulescu e] Iorgulescu, A.: "On BCK algebras - Part V: Classes of examples of finite proper bounded α, β, γ, $\beta\gamma$ algebras and BCK(P) lattices (residuated lattices), with or without conditions (WNM) and (DN)"; submitted.

Figure 8. Reference snapshot from a paper.

4.5. Scenario 5-Commonality in Content

We found that some authors have used very common citation tags. For example, in the reference entry shown in Figure 9 represents a citation-tag "[p]". Here, the contemporary systems will only use the character "P" as a reference tag, as show in Figure 9.



Common character as citation tag

[P] Pásztorné Varga Katalin : A matematikai logika és alkalmazásai, Tankönyvkiadó, Bp, 1986.

Figure 9. Reference snapshot from a paper.

These kinds of citation tags are very sensitive as "P" is common character which may occur many times in the full text of the paper and will mislead the calculation of in-text citation frequencies. The use of common character as a citation tag was one of reasons that caused the overall incorrect marking of in-text citations as shown in Table 1.

These types of problems may be handled by designing proper regular expressions. For example, in the above scenario, the extensive list of regular expression would be as follows: "[P]", "[P,", ",P]", "[P", "P]". "[,P,]".

5. Summary

In-text citation can be used in a number of areas. Therefore, accurate marking of in-text citation is crucial. In this paper, we have presented detailed analysis of in-text citation and some interesting real scenarios explored during manual analysis and verification of in-text citation frequencies. The presented analysis and interesting scenarios will help the researchers to understand the problems of correctly marking of in-text citations automatically. In-text citations are made with the help of citation tag. Different problems have been discussed that are associated with different citation tags such as using only numbers, alphabets and alphanumeric etc., there is a need for a deeper analysis of the content of the paper to better disambiguate between mathematical equation numbers, intervals and the accurate citation-tag. Beside the difficulty of accurate identification of citation tags, there are certain other issues which are related with PDF to text/ XML conversion. The most important are subscript, superscripts and encoding etc. Thus, when devising an automatic solution for in-text citation exploitation, the aforementioned issues must be carefully planned so that maximum accuracy can be achieved.

6. Conclusions

This research focuses on the exploration of in-text citation frequencies in the text of scientific documents. In this paper, we have provided detailed in-text citation analysis on 16,000 citations of an online journal, reported different pattern of citations tags and its in-text citations and presented some interesting real problems that a researcher may confront while exploiting in-text citations. Furthermore, citation tags of inaccurate identification of its in-text citations were divided into three different clusters such as "Numeric", "Alphabetic" and "Single Characters". The "Numeric" and "Alphabetic" clusters were most populated clusters as compared to "Single Character" cluster. Based upon these three types of clusters, different reasons for inaccurate identification of in-text citations were discovered. The frequent errors were due to wrong allotment, mathematical ambiguities, commonality in content and string variations. Finally, we have also highlighted the possible solutions for each problem that will help future systems which focus on the identification of in-text citations in various domains. In future we plan to develop a technique and algorithms to tackle the discussed problems accurately in a systematic way. Moreover, we are planning to build a comprehensive system that can mark various types of the existing in-text citations with sufficient accuracy.

References

[1] Afzal M., Kulathuramaiyer N., Maurer H., and Balke W., "Creating Links into the Future," *the Journal of Universal Computer Science*, vol. 13, no. 9, pp. 1234-1245, 2007.

[2] Afzal M., Maurer H., Balke W., and Kulathuramaiyer N., "Rule Based Autonomous Citation Mining with TIERL," *the Journal of Digital Information Management*, vol. 8, no. 3, pp. 196-204, 2010.

[3] Beel J. and Gipp B., "Google Scholar's Ranking Algorithm: The Impact of Citation Counts (An Empirical Study)," *in Proceedings of the 3rd International Conference on Research Challenges in Information Science*, Fès, Morocco, pp. 439-446, 2009.

[4] Ciancarini P., Iorio A., Nuzzolese A., Peroni S., and Vitali F., "Semantic Annotation of Scholarly Documents and Citations," *in Proceedings of the 13th International Conference of the Italian Association for Artificial Intelligence*, Turin, Italy, pp. 336-347, 2013.

[5] Constantin A., Pettifer S., and Voronkov A., "PDFX: Fully-Automated PDF-to-XML Conversion of Scientific Literature," *in Proceedings of ACM Symposium on Document Engineering*, Florence, Italy, pp. 177-180, 2013.

[6] Garfield E., "Citation Analysis as a Tool in Journal Evaluation," available at: http://www.garfield.library.upenn.edu/essays/V1p527y1962-73.pdf, last visited 2013.

[7] Giles C., Bollacker K., and Lawrence S., "CiteSeer: An Automatic Citation Indexing System," *in Proceedings of the 3rd ACM Conference on Digital Libraries*, Pennsylvania, USA, pp. 89-98, 1998.

[8] Gipp B. and Beel J., "Citation Proximity Analysis (CPA)-A New Approach for Identifying Related Work based on Co-Citation Analysis," *in Proceedings of the 12th International Conference on Scientometrics and Informetrics*, Rio de Janeiro, Brazil, pp. 571-575, 2009.

[9] Goodall A., "Should Top Universities be Led by Top Researchers and are They?: A Citations Analysis," *the Journal of Documentation*, vol. 62 no. 3, pp. 388-411, 2006.

[10] Hirsch J., "An Index to Quantify an Individual's Scientific Research Output," *the Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 46, pp. 16569-16572, 2005.

[11] Iorio A., Nuzzolese A., and Peroni S., "Characterising Citations in Scholarly Documents: the CiTalO Framework," *in Proceedings of Semantic Web: ESWC 2013 Satellite Events*, Montpellier, France, pp. 66-77, 2013.

[12] Iorio A., Nuzzolese A., and Peroni S., "Identifying Functions of Citations with CiTalO," *in Proceedings of Semantic Web: ESWC 2013 Satellite Events*, Montpellier, France, pp. 231-235, 2013.

[13] Iorio A., Nuzzolese A., and Peroni S., "Towards the Automatic Identification of the Nature of Citations," available at: http://ceur-ws.org/Vol-994/paper-06.pdf, last visited 2013.

[14] Liu S. and Chen C., "The Effects of Co-citation Proximity on Co-citation Analysis," *in Proceedings of the 13th Conference of the International Society for Scientometrics and Informetrics*, Durban, South Africa, pp. 474-484, 2011.

[15] Maricic S., Spaventi J., Pavicic L., and Pifat-Mrzljak G., "Citation Context versus the Frequency Counts of Citation Histories," *the Journal of the American Society for Information Science*, vol. 49, no. 6, pp. 530-40, 1998.

[16] Noor S. and Bashir S., "Evaluating Bias in Retrieval Systems for Recall Oriented Documents Retrieval," *the International Arab Journal of Information Technology*, vol. 12, no. 1, pp. 53-59, 2015.

[17] Ritchie A., "Citation Context Analysis for Information Retrieval," *Doctoral Dissertation*, University of Cambridge, 2009.

[18] Shahid A., Afzal M., and Qadir M., "Discovering Semantic Relatedness between Scientific Articles through Citation Frequency," *Australian Journal of Basic and Applied Sciences*, vol. 5, no. 6, pp. 1599-1604, 2011.

[19] Shotton D., Portwin K., Klyne G., and Miles A., "Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article," available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2663789/, last visited 2013.

[20] Teufel S. and Kan M., *Robust Argumentative Zoning for Sensemaking in Scholarly Documents*, Springer Berlin Heidelberg, 2011.

[21] Teufel S., "Citations and Sentiment," available at: http://www.nactem.ac.uk/event_slides/Teufel 291009.pdf, last visited 2013.

Abdul Shahid is a Lecturer in Computer Science at Institute of Information Technology, Kohat University of Science and Technology, Pakistan. Currently, he is pursuing his PhD in computer science from Mohammad Ali Jinnah University Islamabad, Pakistan. His research focuses on recommending relevant documents with the help of in-text citation frequencies and patterns. In this field, he has published number of good quality papers in different international conferences and journals. Beside his research activities, he is a professional software developer and working as consultant for software companies for last six year.

Muhammad Tanvir Afzal earned his masters in computer science (with Gold Medal) from Quaid-i-Azam University, Pakistan, He was awarded PhD with distinction from Graz University of Technology, Austria. He is working as Assistant Professor in the Department of Computer Science at Mohammad Ali Jinnah University, Pakistan, adjunct professor in institute for information systems and computer media at Graz University of Technology, Austria, and Editor-in-Chief for the journal: Journal of universal computer science. He has published more than 60 research papers in well reputed journals and conferences. His research interest includes: digital libraries, semantic web, social web, knowledge management, and sentiment analysis.

Muhammad Abdul Qadir received his PhD degree from University of Surrey GUILDFORD, UK in 1995. He serves as full professor and Dean at Mohammad Ali Jinnah University, Pakistan. He has more than 25 years of experience in industry, academia and management. Currently, he is actively involved in teaching/ R and D and academic management. He is recipient of two research projects of worth more than 55 million rupees. His current research focus is semantic web, multimedia semantics, ontologies, distributed systems and bioinformatics. He has published more than 100 research publications in International Refereed Proceedings and Journals.