HAL open science

Multiple In-text Reference Phenomenon

Marc Bertin, Iana Atanassova

To cite this version:

Marc Bertin, Iana Atanassova. Multiple In-text Reference Phenomenon. Workshop on Bibliometric-enhanced Information Retrieval and NLP for Digital Libraries (BIRNDL2016) co-located with the Joint Conference on Digital Libraries 2016 (JCDL), Jun 2016, Newark, NJ, United States. pp.14-22. hal-01885074

HAL Id: hal-01885074

https://hal.science/hal-01885074v1

Submitted on 1 Oct 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multiple In-text Reference Phenomenon

Marc Bertin[1] and Iana Atanassova[2]

[1] Centre Interuniversitaire de Rercherche sur la Science et la Technologie (CIRST), Université du Québec à Montréal (UQAM), Canada,
bertin.marc@gmail.com
[2] Centre Tesnière, University of Franche-Comté, France,
iana.atanassova@univ-fcomte.fr

Abstract. In this paper we consider sentences that contain Multiple In-text References (MIR) and their position in the rhetorical structure of articles. We carry out the analysis of MIR in a large scale dataset of about 80,000 research articles published by the Public Library of Science in 7 journals. We analyze two major characteristics of MIR: their positions in the IMRaD structure of articles, and the number of in-text references that make up a MIR in the different journals. We show that MIR are rather frequent in all sections of the rhetorical structure. In the Introduction section, sentences containing MIR account for more than half of the sentences with references.

Keywords: Multiple In-text References, Bibliometrics, Citation Analysis, In-text References, Content Citation Analysis, IMRaD structure

1 Introduction

In a scientific article, the frequency of in-text references is highly dependent on the rhetorical structure. For example Bertin et al. [2] study the differences between the four sections of the IMRaD (Introduction, Methods, Results and Discussion) structure in terms of the density of citations.

The phenomenon of multiple in-text references in scientific papers has not yet been studied. In recent works on the rhetorical structure of articles [1–3], the studies are based on the presence of in-text references in sentences but their number in a single sentence is not taken into consideration. Several works exist on a related problem which is the proximity of co-citations in texts [5]. Liu and Chen [9, 10] propose a four level co-citation proximity scheme for the levels of article, section, paragraph and sentence.

Our approach allows to identifies multiple in-text references. In general, the presence of more than one in-text references in the same sentence gives us information about a relative proximity between the works that are cited. Previous studies on in-text references take into account word windows or sentences to study citation contexts. However, some recent works show the importance and the difficulty in identifying citation blocks that are spans of citations that may encompass one or more sentences [7]. Another related question, the recurrence of in-text references or re-citations, has been the object of several studies [6, 15].

We focus on text spans in articles that contain more than one in-text references that appear very close to each other. We consider sentences as a basic textual unit, and we examine sentences containing more than one in-text reference. We call this phenomenon *Multiple In-text References (MIR)*. For example, the following sentence contains a MIR composed of 4 references, where 3 of the references appear in the range *"[74–76]"*:

> *"Indeed, it has long been proposed on thermodynamic grounds that transcription factors would bind at low, nonfunctional levels throughout the genome either via sequence-independent [74–76] or sequence-specific DNA binding [32]."*[3]

In this paper, we present the results on the behavior of MIR and their positions in the IMRaD structure that we obtain by processing a large scale corpus of about 80,000 research articles. The key idea involves identifying the number of in-text references at the level of the sentence. We analyze two major characteristics of MIR. The first one is their positions in the IMRaD structure, and the second one is the number of in-text references that make up a MIR.

2 Method

To address the problem of the identification of MIR and their positions in the rhetorical structure of articles, we have processed a large-scale corpus of articles that follow the IMRaD structure. In fact, during the last decades, IMRaD has imposed itself as a standard rhetorical framework for scientific articles in the experimental sciences.

The objective of our study is to determine the locations where MIR are most likely to appear in the rhetorical structure of scientific articles. We also study the number of in-text references in the sentences.

2.1 Dataset

To perform this study we have analyzed a dataset of seven peer-reviewed academic journals published in Open Access by the Public Library of Science (PLOS). Six of the journals are domain-specific (*PLOS Biology, PLOS Computational Biology, PLOS Genetics, PLOS Medicine, PLOS Neglected Tropical Diseases*) and the 7th is *PLOS ONE*, which is a general journal that covers all fields of science and social sciences. We have processed the entire dataset of about 80,000 research articles in full text published up to September 2013.

The dataset is in the XML JATS format, where the body of the articles consists of sections and paragraphs that are identified as distinct XML elements. The in-text references are also, for the most part, present as XML elements and linked to the corresponding elements in the bibliography of the article.

[3] PLOS Biology, 2008, DOI: 10.1371/journal.pbio.0060027.

2.2 Identification of the IMRaD Structure

To identify the IMRaD structure of articles, we have analyzed all section titles and categorized the sections. All seven journals use similar publication models, where authors are explicitly encouraged to use the IMRaD structure. As a result, more than 97% of the research articles in the corpus contain the four main section types (Introduction, Methods, Results and Discussion), although not always in the same order. While the relative position of a section may have an influence on the number of references found in the section, e.g. a Discussion section that appears immediately after an Introduction section may have more references than a Discussion section that appears at the end of the article, the order in which the sections appear in an article has not been taken into consideration for this study. The detailed results of the analysis of the IMRaD structure for this corpus are presented by Bertin et al. [2].

2.3 Identification and Processing of MIR

In order to identify sentences containing MIR, we need to perform the following steps:

1. Segment all paragraphs into sentences;
2. Identify all in-text references;
3. Consider the number of in-text references in each sentence.

The first step was done by analyzing the punctuation and capitalization of the text in order to identify sentence ends. Our corpus contains a total of 15,852,120 sentences, out of which 3,528,514 (around 22%) contain in-text references.

The second step may seem trivial given that the in-text references are present as elements in the XML tree of the article. In the case of MIR however, this is not always true. When in-text references are in a numeric form, reference ranges are often present in sentences containing MIR. For example, we can consider the following sentence in the corpus with XML markup:

> *"A number of recent studies have used a modification of the picture viewing procedure by substituting pleasant pictures with photographs of loved, familiar faces* <xref ref-type="bibr" rid="pone.0041631-Bartels1">*[16]*</xref> – <xref ref-type="bibr" rid="pone.0041631-Xu2">*[24]*</xref>."[4]

The in-text references in this sentence "*[16]–[24]*" are identified as two *xref* elements that point to the corresponding bibliography items. In reality, the sentence contains 9 different citations: all the works from [16] to [24] are cited; and 7 of these citations are not present in the XML markup. In such cases, we call *implicit in-text references* those in-text references that are part of a range but that are not mentioned by their numbers in the sentence.

In order to identify correctly MIR and their number in sentences it is important to detect in-text reference ranges and implicit references. To do this, we

[4] PLOS ONE, 2012, DOI: 10.1371/journal.pone.0041631.

first examine the context of each *xref* element and identify all possible ranges. Then, we generate the list of implicit references and the links between these references and the bibliography items. A similar method for the processing of in-text reference ranges was used by Bertin et al. [1].

The occurrences of in-text reference ranges are rather numerous in our corpus. We found out that reference ranges are present in 19.19% of all sentences containing MIR and implicit in-text references account for 12.38% of the MIR.

3 Results

We first observe the presence of MIR in the four section types of the IMRaD structure and then we examine the number of the MIR according to the different journals.

3.1 Use of MIR in the IMRaD Structure

Table 1 presents the percentage of sentences containing Multiple In-text References (MIR) among all sentences with in-text references in the four section types of the IMRaD Structure (I-M-R-D). We observe that MIR are present for the most part in the Introduction section where more than half of the sentences with in-text references contain MIR (52.78%). This result is consistent with the observation that the Introduction section often includes a state of the art with a literature review in which MIR are most likely to appear. As for the Methods and Results sections, they have around 25% and 35% of MIR respectively.

	I	M	R	D	*Total*
Sentences with MIR	52.78%	25.05%	35.59%	42.65%	*41.43%*
Sentences without MIR	47.22%	74.95%	64.41%	57.35%	*58.57%*

Table 1: MIR in the IMRaD Structure

The results in this table for the four different section types are not unexpected. In fact, the relative quantity of MIR in the sections follows the overall distribution of references in the IMRaD structure, shown in Bertin et al. [2], where the Methods section contains the smallest number of in-text references, followed by the Results section. The Introduction section, which is also the shortest one on average, contains the highest number of citations.

Table 1 shows also that MIR appear very often: in around 41% of all sentences containing in-text references. This means that in a scientific article, the largest number of in-text references appear in groups of several references situated closely in the textual space, i.e. in the same sentence.

Number of in-text references in MIR	I	M	R	D
2	21.74%	15.67%	19.49%	20.74%
3	11.63%	4.43%	7.31%	9.34%
4	6.25%	1.53%	3.05%	4.36%
5	3.45%	0.68%	1.39%	2.12%
6	2.04%	0.34%	0.70%	1.13%
7	1.22%	0.17%	0.39%	0.64%
8	0.77%	0.10%	0.23%	0.36%
9	0.48%	0.06%	0.14%	0.21%
10	0.32%	0.05%	0.10%	0.13%
11	0.22%	0.03%	0.07%	0.08%
12	0.15%	0.02%	0.05%	0.06%
13	0.10%	0.02%	0.04%	0.04%
14	0.08%	0.01%	0.03%	0.03%
15	0.05%	0.01%	0.03%	0.02%

Table 2: Percentage of sentences with MIR in the IMRaD structure

Table 2 presents the percentage of sentences with MIR of different sizes among all sentences containing citations in each of the four section types. Considering the MIR with 2 elements, we observe that there is little difference between the sections Introduction, Results and Discussion. Then, the differences between the sections increase with the number of in-text references. This means that, while MIR with 2 elements appear almost homogeneously in an article, the MIR with higher number of elements are more and more exclusively reserved to the Introduction section. This phenomenon again, is explained by the presence of the state of the art in the Introduction with a very high concentration of in-text references.

3.2 MIR in the PLOS Journals

Figure 1 presents the relative number of sentences with MIR in each of the journals. The horizontal axis gives the number of in-text references in the same sentence in a logarithmic scale. The vertical axis gives the average number of sentences per article containing MIR in a logarithmic scale.

We observe some differences between the journals in the use of very large MIR (number of elements 20 and above). In fact, the journal PLOS Medicine stands out because it uses MIR with relatively high number of elements.

Table 3 presents the average and the maximal number of elements in MIR observed in the 7 journals. PLOS Medicine has the highest average number of elements in MIR. In fact, articles in this journal tend to be short, but with a very high number of references and many of them appearing in the same sentence. PLOS ONE and PLOS Medicine have very high maximal number of elements in MIR. However, as we can see on figure 1, MIR with a high number of elements in PLOS ONE tend to be less frequent than those in PLOS Medicine.

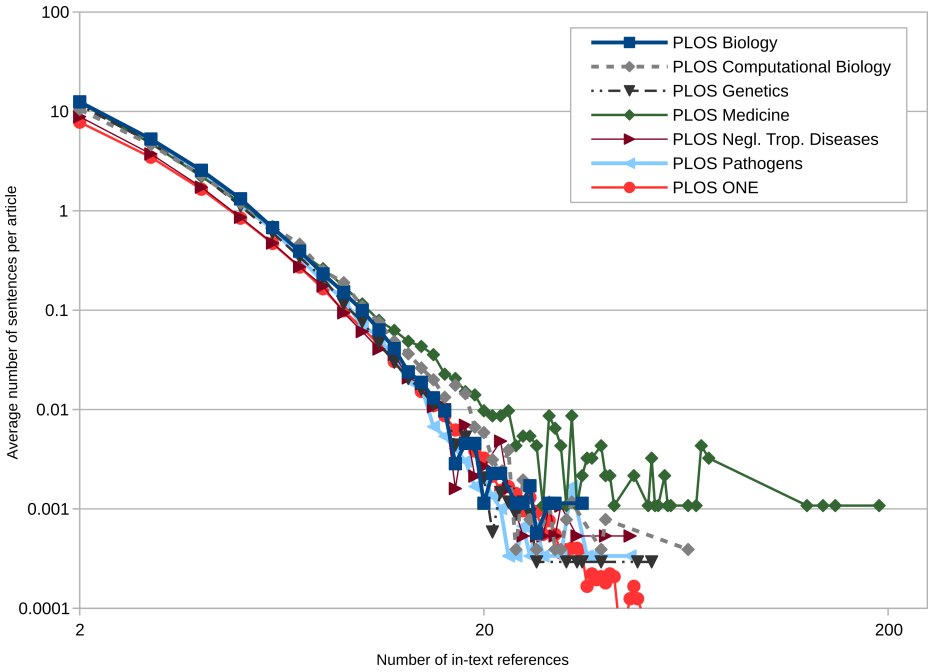

Fig. 1: MIR in the 7 PLOS Journals

Some examples of sentences containing MIR with very high number of In-text References are presented in table 4. In fact, these examples are extracted from articles in the medical domain that are of a specific type: systematic reviews. This kind of articles has for objective to sum up the best available research on a specific topic by collecting and synthesizing the results of other studies that fit pre-specified eligibility criteria. For this reason, we find in these articles sentences that cite a large number of other works. Moreover, as we can see on table 4, these sentences are not necessarily in the Introduction section but appear quite often in the Results and Methods sections.

4 Discussion and Conclusion

We have proposed a study of Multiple In-text References (MIR) in respect of their positions in the rhetorical structure of articles. This study shows the following key points:

- MIR are rather frequent in all sections of articles: 41% of the sentences with citations contain MIR;
- In the Introduction section MIR account for more than half of the sentences containing citations;

Journal	Ave. number of elements in MIR	Standard deviation	Maximal number of elements in MIR
PLOS Biology	3.0496	1.7780	35
PLOS Computational Biology	3.2183	2.0938	64
PLOS Genetics	3.0087	1.7358	52
PLOS Medicine	3.3246	3.8083	190
PLOS Negl. Tropical Diseases	3.0493	1.8583	46
PLOS Pathogens	2.9807	1.6720	46
PLOS ONE	3.1284	2.1384	288

Table 3: MIR in the 7 PLOS Journals: number of elements

Journal	Section	Sentence
PLOS Medicine	R	*The systematic review identified 188 studies that provided prevalence estimates [18,29,36–223].*
PLOS Medicine	M	*Data for calculation of the number of snakebite envenomings were obtained for 46 countries [9–60] while data for calculation of the number of deaths due to snakebite were obtained for 22 countries [9–11,18, 31,47,49,51,56,58–72] by this process.*
PLOS ONE	R	*As a result, 221 unique genes and 4 protein complexes (DNA-PK, HSP70, MRN(95), RAS) were identified from around 200 papers that studied radiation response-related biomarkers [4], [14]–[185].*
PLOS ONE	M	*Details of each study [11]–[113] were entered into a database by one investigator with a 100% re-check.*

Table 4: Examples of sentences with high number of in-text references

- The MIR with two elements are the most homogeneous: they appear quite often in the Introduction, Results and Discussion sections (about 20% of sentences with citations) and in about 15% of sentences with citations in the Methods section;
- There exist sentences with very high number of in-text citations (more than 100). Such sentences are specific to the domain of medicine and the systematic review article type.

In this study, the notion of MIR raises the question of the importance and role of MIR in scientific articles. We show the behavior and location of sentences that contain MIR. The implications of this study are relevant from the perspective of networks as bibliographic coupling [8], clustering [14] and co-citation [13], but also for the analysis of the functions of citations. Furthermore, many applications can benefit from the differentiation between single and multiple references such as automatic summarization [4, 12] or automatic generation of surveys [11]. More generally, the distribution of MIR along the text progression has impor-

tant implications for understanding the contexts of citations. For example, we can consider the following sentence:

> *"Previous attempts to apply functional genomics methods to address these questions used various approaches, including* ***DNA microarrays*** *(Hayward et al. 2000; Ben Mamoun et al. 2001; Le Roch et al. 2002),* ***serial analysis of gene expression*** *(Patankar et al. 2001), and* ***mass spectrometry*** *(Florens et al. 2002; Lasonder et al. 2002) on a limited number of samples from different developmental stages."*[5]

In this sentence, there are 3 groups of in-text references and each group is characterized by a noun group that identifies topics related to the in-text references. The automatic identification of these topics will allow to assign them to each of the references.

This example shows that work at the level of sentences is not enough if we want to obtain fine and accurate results for content citation analysis. The observations of this study suggest the presence of MIR implies the existence of features such as topics, keywords, methods, etc. that are common to all works cited in the MIR group. This means that by examining the text content of such sentences one can obtain information on the topics that are shared by the group of cited works.

5 Acknowledgments

We thank Benoit Macaluso of the Observatoire des Sciences et des Technologies (OST), Montreal, Canada, for harvesting and providing the PLOS data set.

References

1. Bertin, M., Atanassova, I., Larivière, V., Gingras, Y.: The distribution of references in scientific papers: an analysis of the imrad structure. In: 14th International Society of Scientometrics and Informatics Conference. International Society for Scientometrics and Infometrics, Vienna, Austria (July 15-19 2013)
2. Bertin, M., Atanassova, I., Larivire, V., Gingras, Y.: The invariant distribution of references in scientific papers. Journal of the Association for Information Science and Technology 67(1), 164177 (January 2016)
3. Ding, Y., Liu, X., Guo, C., Cronin, B.: The distribution of references across texts: Some implications for citation analysis. Journal of Informetrics 7(3), 583–592 (2013)
4. Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., Radev, D.: Blind men and elephants: What do citation summaries tell us about a research article? Journal of the American Society for Information Science and Technology 59(1), 51–62 (2008)

[5] PLOS Biology, 2003, DOI: 10.1371/journal.pbio.0000005.

5. Gipp, B., Beel, J.: Citation Proximity Analysis (CPA) A new approach for identifying related work based on Co-Citation Analysis. In: Larsen, B., Leta, J. (eds.) 12th International Conference on Scientometrics and Informetrics. vol. 2, pp. 571–575. International Society for Scientometrics and Informetrics, Rio de Janeiro, Brazil (July 14-17 2009)
6. Hu, Z., Chen, C., Liu, Z.: The recurrence of citations within a scientific article. In: Salah, A., Tonta, A., Akdag Salah, C., Sugimoto, U.A. (eds.) 15th International Society of Scientometrics and Informetrics Conference. International Society for Scientometrics and Infometrics, Bogazii University Printhouse, Istanbul, Turkey (June 29 to July 3 2015)
7. Kaplan, D., Tokunaga, T., Teufel, S.: Citation block determination using textual coherence. Journal of Information Processing 24(3), 540–553 (May 2016)
8. Kessler, M.M.: Bibliographic coupling between scientific papers. American documentation 14(1), 10–25 (1963)
9. Liu, S., Chen, C.: The proximity of co-citation. Scientometrics 91(2), 495–511 (2011)
10. Liu, S., Chen, C.: The Effects of Co-citation Proximity on Co-citation Analysis. In: 13th Conference of the International Society for Scientometrics and Informetrics. vol. 1 and 2, pp. 474–484. International Society for Scientometrics and Infometrics, Durban, South Africa (July 4-7 2011)
11. Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D., Zajic, D.: Using citations to generate surveys of scientific paradigms. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 584–592. Association for Computational Linguistics, Boulder, Colorado, USA (May 31 to June 5 2009)
12. Qazvinian, V., Radev, D.R.: Scientific paper summarization using citation summary networks. In: Proceedings of the 22nd International Conference on Computational Linguistics. vol. 1, pp. 689–696. Association for Computational Linguistics, Manchester, UK (Aug 18-22 2008)
13. Small, H.: Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for information Science 24(4), 265–269 (1973)
14. Small, H., Sweeney, E., Greenlee, E.: Clustering the science citation index using co-citations. ii. mapping science. Scientometrics 8(5-6), 321–340 (1985)
15. Zhao, D., Strotmann, A.: Re-citation analysis: Promising for research evaluation, knowledge network analysis, knowledge representation and information retrieval? In: 15th International Society of Scientometrics and Informatics Conference. pp. 1061–1065. International Society for Scientometrics and Infometrics, Bogazii University Printhouse, Istanbul, Turkey (June 29 to July 3 2015)