Potential Predictibility of References in the Identification of Derivative Articles from Doctoral Theses

Mercedes Echeverria ***David Stuart*** ***Tobias Blanke***

Abstract

This paper reports the results obtained on the predictability of references for the identification of derivative articles from doctoral theses, based on a sample of 68 medical theses and 334 articles published by the same theses authors. The study performs an analysis of the common references shared by theses and articles through a text similarity approach. A textual similarity comparison is carried out with the discursive sections of articles (Introduction, Methodology, Results and Discussion) based on the full-text of theses and articles. The results suggest that the Reference section has a high sensitivity to detect true positives cases and a low specificity to identify negative cases, corresponding to a high recall a low precision in the detection of derivative articles.

Keywords: Derivative Articles, Doctoral Theses, Cluster Analysis Methodology

1. Introduction

A doctoral thesis consists in an original project of research whose results contribute to new knowledge in a discipline. Based on the premise that the final product of research is the publication of findings in peer-reviewed articles, the objective of this study is to explore the potential predictability of the references to detect derivative articles emerging from PhD theses and determine their typological features.

The study concentrates on medical theses, which represent an important source of publications in medical literature. A study of the Observatorie des Sciences et Techniques (2002) estimated that the research activity of PhD students represented 10-20% of indexed academic research within Scientific, Technical and Medical (STM) publications. In a recent study Larivière (2012) reported that the contributions from PhDs accounted for about a third of publications output in natural and medical sciences.

This paper explores the common references shared between theses and articles published by the theses authors through a text similarity approach. This same approach is used to rank the textual similarity of the discursive sections of articles (Introduction, Methodology, Results and Discussion) based on the full-text of theses and articles. The aim of this study is to assess the potential capacity of references to detect derivative articles and subsequently to compare this data to other discursive sections of articles.

For the purpose of this work, we define derivative works based on three factors that need to come together. A scientific article will be derivative if there exists a textual similarity between thesis and article, if thesis and article share authorship, and finally if thesis and article are published in close temporal proximity.

Text similarity was detected using the anti-plagiarism tool Turnitin, a commercial software intended for verifying the originality of scholarly content. The criteria to validate Turnitin as instrument of analysis

were: dimensionality to process large quantities of text, as in our study a medical thesis contains 72,433 words as average; capacity to measure the degree of the textual similarity between two documents, Turnitin highlights the text location of matches, and feasibility to operate intra-corpal, where the source and copy takes part of the same corpus of the database. In this respect, Turnitin works as a text-matching software.

Turnitin operates on the basis of creating digital fingerprints, which are used to compare documents to each other. The detection of similarity is dependent of several variables such as long strings of consecutive words, small variation of the order of words and small changes consecutive strings within a fragment, all of them common to bibliographic description formats. Turnitin processes the references at level of a textual string rather than as references to a specific document. The text similarities detected in references may be compared to bibliographic coupling, introduced by Kessler (1963), where two documents are bibliographically coupled if they reference the same document. However, although the two models have significant similarities the concepts are not exactly the same. The correlation between these units of analysis, bibliographic coupling and reference text similarity may be categorized in three ways: (i) complete correlation between textual similarity and bibliographic coupling, when both reference a common third document with the same text; (ii) bibliographic coupling and partial text similarity, when two documents cite the same third work but the text similarity is not complete, due to variations of the form of references; (iii) partial text similarity but not bibliographic coupling, when it there is a degree of text similarity between two references but the document cited is not the same. Whilst bibliographic coupling has been more extensively researched than reference text similarity, reference text similarity is an important corresponding methodology because it allows the comparison of references in documents that have not had their references indexed (e.g., theses).

3. Background

Many bibliometric studies have investigated citations linkages and text minning similarity to identify to relations between publications, discover emerging research topics and map scientific fields.

Different approaches have investigated to what extent text mining and bibliometric methods can supplement each other and whether they can improve individual approaches. In a study by Ahlgren and Colliander (2009) the similarity of documents was measured comparing citations to text-based approach. They found that citation-only methods performed worse that text-only methods.

Boyack and Klavans (2010) compared three pured citation-based techniques, direct citations, bibliographic coupling, and co-citation analysis and a citation text hybrid approach with the aim of selecting the network that could best represent the research front in biomedicine. They found that a citations-text hybrid approach outperforms other approaches. Several studies have confirmed that the combination of full-text analysis and bibliometric methods improves the individual methods (Glenisson et al., 2005; Janssens et al., 2006; Ahlgren and Colliander, 2009). The outcomes showed the advantage of the hybrid approach improves upon the bibliometric methods in all respects. However, other studies have examined the validity of these methods identifying limitations, Zitt et al. (2011) compared the citation-based and word-based on a large-size docu-

ment sets in the nanoscience field. They found that the convergence of these approaches could yield quite different outcomes and cannot be substituted each other. Yan and Ding (2012) pointed also out that the hybrid approach could construct heterogeneous scholarly networks related to each other.

The present analysis pursues three aims:

- To assess quantitatively the capacity of the Reference section to identify derivative articles
- To compare the potential of Reference section to other discursive sections to detect derivative articles
- To describe the typological features of derivative articles

4. Data and Methodology

The dataset comprised 68 biomedical theses published in Open Access between 2007 to 2012 and 334 articles published in peer-reviewed journals by the same theses authors.

The first step consisted in processing the theses and articles in Turnitin with the objective of gathering statistical data of textual similarity. Turnitin generates its similarity index as a percentage based on a summary of matching similar text found in the document submitted, in this case, the articles against the document target, the theses.

Secondly, we wanted to know the distribution of text similarity among all sections in the articles (Introduction, Methodology, Results and Discussion and References) with the purpose of analyzing the levels of similarity of different sections. For that, we computed the data of matches of similarity of each section.

Given that the starting point was to cluster the articles according to their structure, a proximity matrix was used to measure the distance between all pairs of clusters. This approach would provide insights about which sections produce the highest indexes of similarity, the values of proximity and distance between the different sections, which sections are closest and how they are related to each other. To understand how to perform the clusters of different sections a dendrogram was constructed using the average linkage among groups. This dendrogram displays how strongly the individual sections are correlated, based in their degree of similarity.

The final objective of research was to identify the optimal cut-off within the sections of articles with high levels of similarity in order to discriminate derivative from non-derivative works. To this end, we used the Receiver Operating Characteristic (ROC) curve. The ROC curve is a two-dimensional graph that visually depicts the full picture of trade-offs between the sensitivity (proportion of true positives) and 1-specificity (proportion of false positives) across a series of cut-off points in order to identify correctly the optimal threshold point to predict the derivative works.

5. Results

Seven different clusters were generated according to the sections of the articles (Title, Abstract, Introduction, Methodology, Results, Discussion and References). An Excel spreadsheet was used to compute the distribution of matches along sections (See Appendix).

With the aim of analyzing the distance of the sections of articles a matrix was constructed (Table 1), the specifications of the matrix were Euclidean dis-

tance squared and the initial values were transformed by the type, range 0-1. The matrix showed that the clusters with closed distances were Discussion and Results (3.459) followed by Introduction and Discussion (3.825) and Methodology and References (4.860).

Table 1: Distance Matrix between Clusters

Proximity Matrix							
Case	Matrix File Input						
	Title	Abstract	Introduction	Methodology	Results	Discussion	References
Title	,000	7,887	10,004	10,627	8,738	8,265	9,946
Abstract	7,887	,000	5,631	7,686	6,396	5,168	8,422
Introduction	10,004	5,631	,000	6,591	5,626	3,825	6,658
Methodology	10,627	7,686	6,591	,000	5,499	5,000	4,860
Results	8,738	6,396	5,626	5,499	,000	3,459	5,408
Discussion	8,265	5,168	3,825	5,000	3,459	,000	5,347
References	9,946	8,422	6,658	4,860	5,408	5,347	,000

The dendogram (Fig. 1) provides the rescaled distance how clusters are combined..

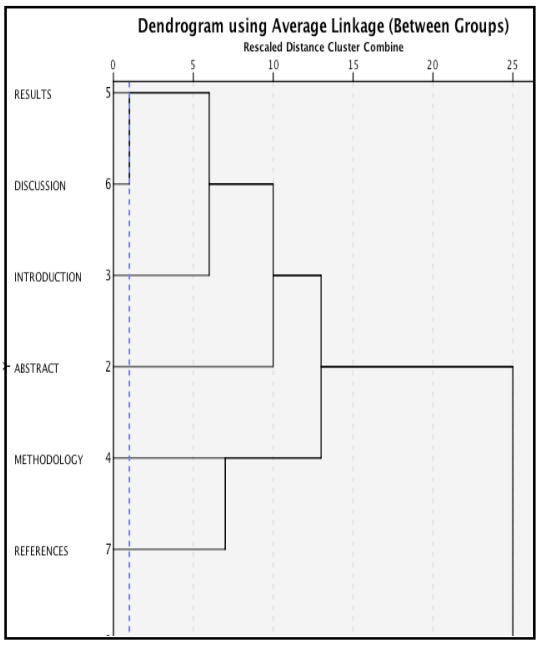


Figure 1: Hierarchical cluster dendrogram of the distances between the sections of articles

The dendrogram consists of 6 clusters. The most significant is (Discussion – Results). This first agglomeration would indicate the linkage between these discourse sections. These sections are interdependent, Discussion extrapolates and describes the Results. From a rhetorical perspective Discussion explains the significance of Results and emphasizes their accuracy and consistency. The other clusters (Introduction – Discussion – Results) and (References – Methodology) represent small distances between them, the discrimination in small distances is not significant. They are statistically blurred groups, heavily correlated, where it is complex identify categories. In fact, small variations in the data distances could yield different conglomerates, whereas the cluster (Results – Discussion) would remain stable.

In order to identify the optimal cut-off point within the sections of articles with high levels of similarity ROC curve test was used. The study was designed in two phases. The first one consisted in defining the decision rules to construct the 'Gold Standard', value of reference necessary to calculate the sensitivity and specificity.

The second phase checked the capacity of the test in terms of differentiating derivative from non-derivative articles.

The decision rules based on external evidences for constructing the Gold Standard were:

- Theses authors' statements confirming the relation between theses and articles
- Complete textual similarity between the articles titles and theses chapters
- Articles bound in the doctoral theses

Table 2: External Criteria used for Decision Rules of Gold Standard

Authors' statements	Textual titles similarity (Articles/Theses)	Articles bound in theses	Total
18	26	15	59

Corpus analysed: 334 articles

Gold Standard: 46 articles.

More than one criterion was common to 13 articles.

The ROC curve estimated the optimal threshold point to predict derivative works (Fig. 2)

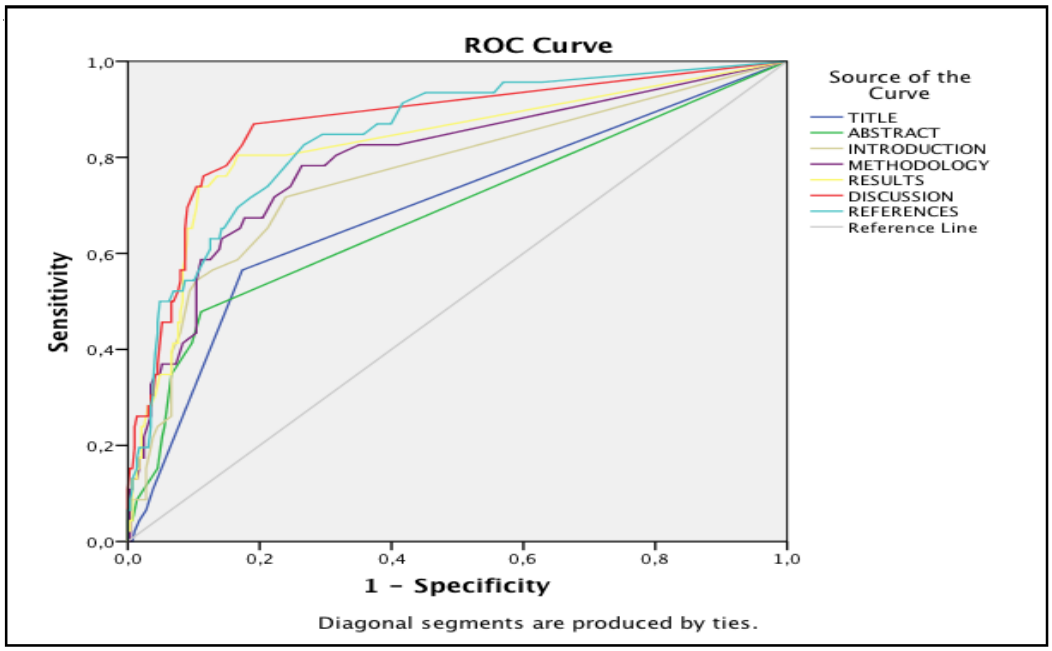


Figure 2: ROC curve of derivative articles. Area under curve

The area under the curve (AUC) corresponds to the probability of identifying derivative and non-derivative articles within two dichotomous variables. The larger is AUC the better is overall performance the test.

The results showed (Table 3) that the greater AUC corresponded to the Discussion section with 0,869 (95% CI 0.807 -0.931), which suggests that Discussion is very predictable with respect the identification of derivative articles, followed by the Reference section with a surface under curve of 0,846 (95% CI 0,785-0,907).

Table 3: Matrix of Distances of the Sections of Articles

Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
TITLE	,694	,046	,000	,605	,784
ABSTRACT	,685	,048	,000	,591	,779
INTRODUCTION	,764	,042	,000	,682	,847
METHODOLOGY	,796	,039	,000	,719	,873
RESULTS	,829	,038	,000	,755	,903
DISCUSSION	,869	,032	,000	,807	,931
REFERENCES	,846	,031	,000	,785	,907

The test result variable(s): TITLE, ABSTRACT, INTRODUCTION, METHODOLOGY, RESULTS, DISCUSSION, REFERENCES has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

The method used to calculate the optimal threshold point was the Youden Index (Kumar and Indrayan, 2011), obtained by deducting 1 from the sum of test's sensitivity and specificity, expressed not as percentage but as a part of a whole number.

The cut-point calculated in Discussion was 3.50 (sensitivity 0.761). The cut-point calculated in References was 17.50 (sensitivity 0.826) .

Table 4: Coordinates of the ROC curve. Discussion

Test Result Variable(s)	Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity	Specificity	Sensitivity + Specificity
Discussion	3.50	0.761	0.115	0.885	1.646
References	17.50	0.826	0.267	0.733	1.559

Cut-Point True Positives [TP] False Positives [FP] True Negatives [TN]

The difference in the results of the ROC curve using the 3.50 cut point and the 17.50 cut point are showed in the table below:

Table 5: Values of the Roc curve, Discussion and References

Articles	Cut Point	Gold Standard (+)(46)	Gold Standard (-)(288)	Corpus 334
Derivatives	3.50 (Discussion)	35 [TP]	33 [FP]	68
	17.50 (References)	38 [TP]	76 [FP]	114
Non-derivatives	3.50 (Discussion)	11 [FN]	255 [TN]	266
	17.50 (References)	8 [FN]	212 [TN]	220

Statistical comparison of results of the ROC curve between Discussion and References section.

1. The number of articles identified as derivative:
   a) Discussion section: 68 out 334 (20.3%)
   b) References section: 114 out 334 (34.13%)
2) Textual similarity Index percentage:
   a) Discussion section: 36.53%
   b) References section: 24.65%
3) Textual similarity Index median:
   a) Discussion section: 34.50
   b) Reference section: 18.00
4) Thesis' author position as first author of the articles
   a) Discussion section: 86.76%
   b) Reference section: 64.03%
5) Co-authorship and thesis supervisors
   a) Discussion section: 68 (100%), supervisors collaborated as co-authors in the articles within this range:

   1/2= 19/68 articles

   1/3= 1/68 articles

   2/2= 25/68 articles

   2/3= 6/68 articles

   3/3= 1/68 articles

a) Reference section: 107 (93,85%) supervisors collaborated as co-authors in the articles within this range:

   1/1= 24/107 articles

   1/2= 41/107 articles

   1/3= 4/107 articles

   2/2= 27/107 articles

   2/3= 7/107articles

   3/3= 4/107 articles

6) Number of authors by article:
   a) Discussion section: 4.68 authors/articles (SD 2.39)
   b) References section: 6.03 authors/articles (SD 3.33)

7) Time differential of articles regarding thesis in Open Access (OA) and article publication online:

   a) Discussion section: 63/68 (92.64%) articles were published before thesis would be in OA.

   b) References section: 86/114 (75.43%) articles were published before thesis would be in OA.

Furthermore, it was observed by means a plot of the univariate analysis of variance (ANOVA procedure) the variation of common references over time. Results showed that the highest similarity in references is produced between one or two years before the thesis is published in OA, falling sharply the first year that a thesis is available in OA, as shown in Figure 3.

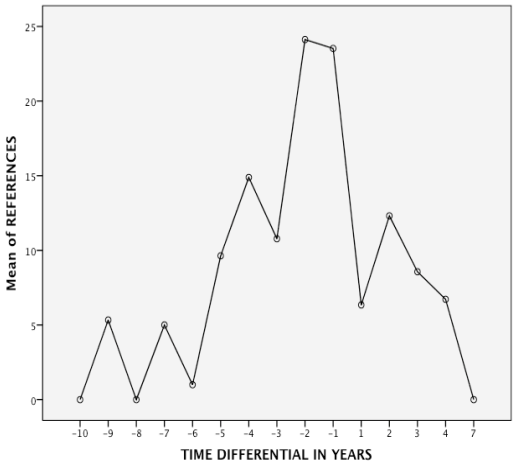
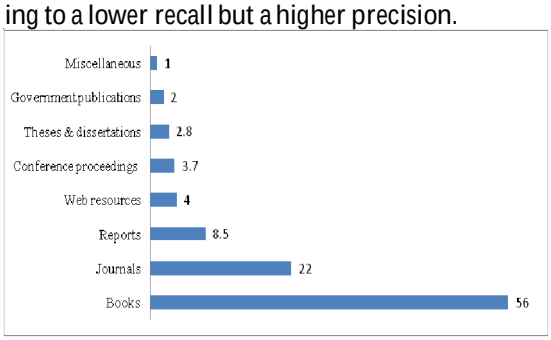


Figure 3. Evolution over time the common references of theses and articles.

6. Discussion

The analysis showed that references have a high sensitivity or capacity to detect correctly true positive cases [TP] (0.826 = 38 cases) and a low specificity to identify correctly negative cases [TN] (0.733 = 212) corresponding to a high recall and low precision. In contrast, the text similarity approach showed that the Discussion has a lower sensitivity [TP] (0.761 = 35 cases) but a higher specificity to identify correctly negative cases [TN] (0.885 = 255 cases), corresponding to a lower recall but a higher precision.



the standardized format of references in publications, whereas the likelihood of processing false positives (high text similarity between references to different papers) may be relatively higher due to very similar vocabulary used in local environments. The differences between text similarity and bibliographic coupling are an area for further exploration in the future.

All indicators showed that the Discussion section is more sensitive and reliable than references in detecting derivative articles and that the Discussion section outperforms the reference sections in all respects. Additionally, these results allowed us to know the typological characteristics of derivative articles:

- High textual similarity between thesis and article (Discussion, 36.53%)
- The thesis's author position of first author (Discussion, 86.76%)
- Presence of the supervisors as co-authors of articles. (Discussion, 100%)
- Low number of authors by article (Discussion, 4.68)
- Articles published before theses in OA (Discussion, 92.64%).

In terms of textual similarity, the high similarity of the Discussion section determines a strong correlation between theses and derivative articles. This finding was already pointed out by Echeverría et al. (2015) and it is consistent with Sun et al. (2010) "The probability of high abstract similarity given similar Results/Discussion sections is significantlyhigher than the probabilities of high Abstract similarity given similar Methods section because the novelty of research articles is typically in its Results/Discussion sections".

Regarding the authorship, the order of authors the study showed the prevalence of the thesis's author as the first author in (86,76%) of derivative articles. This result reflects the general premise by which the name of the principal investigator is almost always mentioned first (Subramanyan, 2983). However, other studies conducted on outputs of theses showed different results (Arriola-Quiroz et al. 2010; Dhaliwal et al., 2010; Diez et al. 2000). These investigations confirm that the order by which authors are listed on a paper is a complex topic and one of the least standardized ones (Jones and McLellan 2000). As it is indicated by the ICMJE (2010) "the decision of the order of authorship is to the coauthors".

The participation of supervisors has proven to be an essential component of derivative articles. The participation of PhD students in research teams was already noted by Larivière (2012) as a determining factor for science students. According to this study, it seems that the integration of PhD students occurs at intra team level, with a high rate of collaboration (67.06%) between supervisors and of PhDs students. This evidence is consistent with the findings of Pole, mentioned by Larivière (2012) " [even] if we do not have any information of the link between the student and other authors, it could be expected, that those co-authors were supervisors and mentors".

Regarding the number of authors, there are discernible differences between derivative articles (4.67 mean) to the average number of authors per article in biomedical journals 6.9 (Weeks et al. 2004) or 6.69 (Costas et al., 2011). These differences could be linked to two factors: the authorship credit, determined by the contribution of the thesis's author in the production of the article and the type of collaboration of PhD students at the intra team level, governed by personal interactions.

Another general feature that emerges from our analysis was the high correlation between the date of publication of the derivative articles and the date of the thesis publication in OA (92,64%). This characteristic appears to be linked to several factors: reliability and validity of peer-review, which ensures rigorous evaluation and increases the quality of thesis, the entrance of PhDs to academic careers and institutional policies for doctoral degrees, and perceived risks of PhD students to publish in journals of high impact, due to potential conflict of interest to journal publishers (Stanton and Liew, 2011).

Regarding the comparison of results obtained, the Reference section revealed a high overlap with the Discussion section (Appendix). The distribution was:

-59 out of 68 articles obtained of the Discussion section were in the list of articles of the Reference section, with a similarity index of 37.78%.

-55 out 114 articles of the list of the References section were not counted in the list of the Discussion section, while their similarity index was of 10.69%.

7. Conclusions

Overall, we can conclude that derivative articles share content and references, that the Discussion section is more sensitive than the Reference section

to detect derivative articles and that references have a high recall but a low precision in detecting derivative articles.

The concepts references text similarity and bibliographic coupling are not exactly the same. The potential relationship between bibliographic coupling and text similarity may be the subject of future studies. Whilst bibliographic coupling has been more extensively researched, reference text similarity may been as an important corresponding methodology, because it allows to compare references in documents that have not had their references indexed, as doctoral theses and articles.

References

1. AHLGREN, Per, COLLIANDER, Cristian. (2009). Document–document similarity approaches and science mapping: Experimental comparison of five approaches. Journal of Informetrics, Vol. 3 (1), pp. 49-63.
2. ARRIOLA-QUIROZ, Isaias [et al.], (2010). Characteristics and publication patterns of theses from a Peruvian medical school. Health Information & Libraries Journal, Vol. 27(2), pp.148-154.
3. BOYACK, Kevin W.,KLAVANS, Richard. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? Journal of the American Society for Information Science and Technology, Vol. 61(12), pp. 2389-2404.
4. COSTAS, Rodrigo, BORDONS, María. (2011). Do age and professional rank influence the order of authorship in scientific publications? Some evidence from a micro-level perspective. Scientometrics, Vol. 88(1), pp.145-161.
5. DHALIWAL, Upreed, SINGH, Navjeevan, BHATIA, Arati. (2010). Masters theses from a university medical college: Publication in indexed scientific journals. Indian Journal of Ophthalmology, Vol. 58(2), pp. 101-104.
6. DIEZ, Claudius, ARKENAU, Cord, MEYER-WENTRUP, Frierike. (2000). The German medical dissertation-time to change?. Academic Medicine, Vol. 75(8), pp. 861-863.
7. ECHEVERRÍA, Mercedes, STUART, David, BLANKE, Tobias. (2015). Medical theses and derivative articles: dissemination of content and publication patters. Scientometrics, Vol. 102 (1), pp. 559-586
8. GLENISSON, Patrick, GLÄNZEL, Wolfgang, PERSSON, Olle. Combining full-text analysis and bibliometric indicators. A pilot study. Scientometrics, 2005, vol. 63, no 1, pp. 163-180.
9. INTERNATIONAL COMMITTEE OF MEDICAL JOURNAL EDITORS. (2010) Uniform requirements for manuscripts submitted to biomedical journals. Available at http://www.icmje.org/urm_full.pdf. (Accessed August 13, 2013).
10. JANSSENS, Frizo [et al.]. (2006) "Integration of textual content and link information for accurate clustering of science fields." Proceedings of the I International Conference on Multidisciplinary Information Sciences & Technologies (InSciT2006). Current Research in Information Sciences and Technologies. Volume I. 2006. pp. 615-619.
11. JONES, Anne Hudson. MCLELLAN, Faith. (ed.) (2000). Ethical issues in biomedical publication. Baltimore: Jonhs Hopkins University.

12. KESSLER, Maxwell Mirton. (1963) Bibliographic coupling between scientific papers. American documentation, vol. 14, no 1, p. 10-25.

13. KUMAR, Rajeev, INDRAYAN, Abdgaya (2011). Receive operating characteristic (ROC) curve for medical researchers. Indian pediatrics, Vol. 48(4), pp. 277-287.

14. LARIVIÈRE, Vincent. (2012). On the shoulders of students? The contribution of PhD students to the advancement of knowledge. Scientometrics, Vol. 90 (2), pp. 463-481.

15. OBSERVATOIRE DES SCIENCES AND TECHNIQUES (2002). Indicateurs bibliometriques des institutions publiques de recherché. Mentioned by Paillassard, P. Schöpfel, J., Stock, C. (2007). Dissemitation and preservation of French print and electronic theses. The Grey Journal, Vol. 3 (2), pp. 77-93.

16. SHIBATA, Naoki [et al.] (2009). Comparative study on methods of detecting research fronts using different types of citation. Journal of the American Society for Information Science and Technology, Vol. 60(3), pp. 571-580.

17. STANTON, Kate V., LIEW, Chern Li (2011) Risks, Benefits and Revelations: An Exploratory Study of Doctoral Students' Perceptions of Open Access Theses in Institutional Repositories. Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation. Springer Berlin Heidelberg, pp. 182-191.

18. SUBRAMANYAM, Krishnappa, (1983). Bibliometric studies of research collaboration: A review. Journal of Information Science, Vol. 6(1), 33-38.

19. SUN, Zhaohui [et al.] (2010). Systematic characterizations of text similarity in full text biomedical publications. PloS one, Vol. 5(9), pp. e12704.

20. WEEKS, William.B., WALLACE, Amy.E., KIMBERLY, BC Surott. (2004). Changes in authorship patterns in prestigious US medical journals. Social Science & Medicine, Vol. 59, (9), pp. 1949-1954.

21. YAN, Erjia, DING, Ying. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. Journal of the American Society for Information Science and Technology, Vol. 63(7), pp. 1313-1326.

22. ZITT, Michel, ALAIN Lelu, BASSECOULARD, Elise (2011). Hybrid citation-word representations in science mapping: Portolan charts of research fields?.. Journal of the American Society for Information Science and Technology, Vol. 62 (1), pp. 19-39.

About Authors

Mercedes Echeverria, Librarian, M.A. and PhD student, Library of the Autonomous University of Madrid.
Email: mercedes.echeverria@uam.es

Dr. David Stuart, Research Fellow, Centre for e-Research, King's College London
Email: david.stuart@kcl.ac.uk

Dr. Tobias Blanke, Senior Lecturer, Centre for e-Research, King's College London
Email: tobias.blanke@kcl.ac.uk