Spring May 4, 2017

Automatic selection of references for the creation of a biomedical literature review using citation mapping

Houcemeddine Turki

Automatic selection of references for the creation of a biomedical literature review using citation mapping

Houcemeddine Turki
Faculty of Medicine of Sfax
University of Sfax
Sfax, Tunisia
turkiabdelwaheb@hotmail.fr

***Abstract*—Biomedical literature review are important research items that significantly contribute to the amelioration of the scientific awareness of physicians, pharmacists, health professionals as well as biologists. That is why several scientists had interest in the last few years in how to automate the process of the creation of such research publications. Unfortunately, the latest advances in citation analysis excepting the co-citation network analysis are not used in this automation process as it should be. That is why I define in this work a full workflow of how to use a combination of direct citation network analysis, basic Scientometrics and co-citation network analysis to choose the most relevant references to be considered in a given literature review from an initial set of papers.**

Keywords—Literature review, systematic review, narrative review, automation, citation analysis, citation network, co-citation network, citation mapping

I. INTRODUCTION

Known as a written work that gathers and discusses multiple research studies or papers about a given topic, literature reviews have always been useful to increase and update the knowledge of scientists about the different research topics and mainly about the biomedical ones and consequently to contribute to the amelioration of the general level of research in all aspects (involving molecular, genetic, and social ones) of medicine, of biology and of other disciplines [1]. Consequently, they are receiving more interest from researchers than the other types of journal items when finding references or trying to understand a research topic they work on [2].

As this type of researches are judged as useful by physicians, biologists and health specialist, as the manual update of literature reviews can take time, and as biomedical knowledge is developing every day, biomedical researchers had tried in the last two decades to automate the creation of this type of publication [1, 3]. In fact, quite all the defined steps to create a biomedical literature review (systematic or narrative one) that are shown in Fig. 1 were partly or absolutely automated [1, 4, 5, 6, 7].

Particularly, significant researches have interested in these years in how to drop irrelevant or useless references from the initial set of references considered for a literature review (that are mainly automatically retrieved from Medline, the largest medical research database) [1, 6, 8, 9].

It is true that most of these works are based on applying the Natural Language Processing techniques (like co-word analysis) on the titles, the abstracts and the keywords of the references [10, 8, 11].

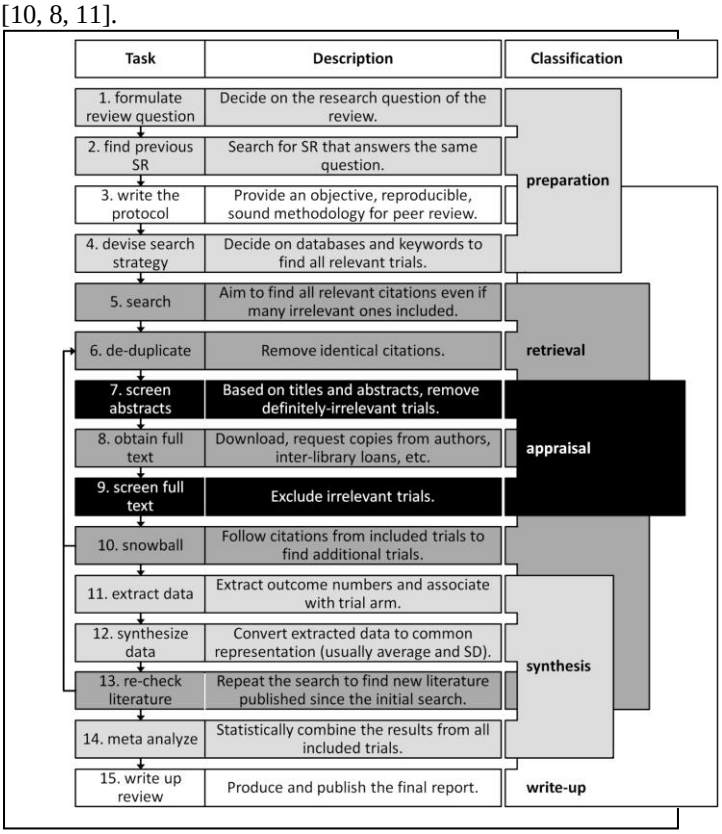


Fig. 1. Method of the creation of a biomedical systematic review as defined by PRISMA Group in 2009 [1, 12]. Only step 12 and step 14 slightly differ in the method for the creation of a biomedical narrative review [13, 14, 15].

However, an important number of research papers had also interested in how citation analysis (mainly citation mapping) can be useful in eliminating irrelevant or useless papers from the initial set of considered references for a given literature review.

As explained in [16], when two articles are cited together in many papers, it is likely that these papers are similar. Consequently, it is useless to cite both articles in a literature review. That is why the papers with the best weight in the co-

citation network of the topic are probably the main papers that report most of the findings about topic [17].

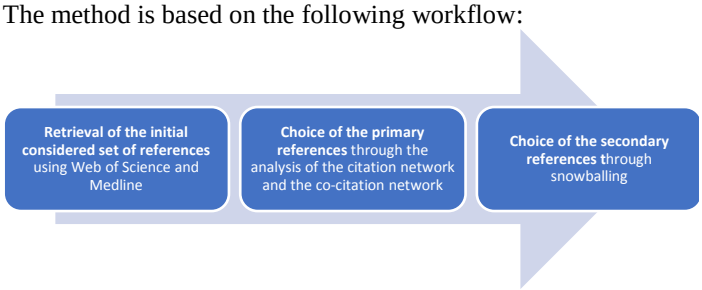
Similarly, papers that are used as references in many papers about the topic are probably the main ones that were used to build all the findings about the topic [18, 19]. That is why the papers with the best weight in the direct citation network of the topic are probably the main papers that explain most of the findings and the historical development of the topic [17, 18].

As the papers having the best weights in direct citation networks are not the same as the ones having the best weights in co-citation networks, it is advised that these two methods are integrated to select the considered papers for a literature review [17]. Unfortunately, in practice, existing literature reviews based on citation mapping use either direct citation network analysis [20, 21] or co-citation network analysis [22, 23, 24].

In this research paper, I define a method that integrate direct citation network analysis and co-citation network analysis to select the references of a given literature review from a list of MeSH-based Medline query results.

II. Proposed Method

The method is based on the following workflow:



A. *Retrieval of the initial considered set of references*

As seen in the introduction, the research question will be converted to a MeSH-based query before proceeding to using a specialized software or directly by searching MeSH keyword database available in https://www.ncbi.nlm.nih.gov/mesh/ [1, 6]. Further details about doing this can be found in [1], [6] and [25]. An explanatory example for this step is provided in Fig. 2.

Use of drugs to cure osteoporosis in Europe
becomes
"Osteoporosis/drug therapy"[Mesh] AND "Europe"[Mesh]

Fig. 2. Example of the conversion of a research question to a MeSH based query

After the MeSH-based query is formulated, I use Web of Science, a paid website managed by Clarivate analytics and dedicated to the citation and bibliometric analysis of scientific literature, to retrieve the search results of the query from Medline [26] (This method to retrieve data from Medline has first been reported by Loet Leydesdorff et al. between 2013 and 2015 [27, 28]). To do that, I need to log in to Web of Science website, choose Medline as a database (Fig. 3), enter the query and choose "MeSH headings" as the type of the used keywords (Fig. 4). After doing these steps and clicking on Search button, I find the search results (Fig. 5).

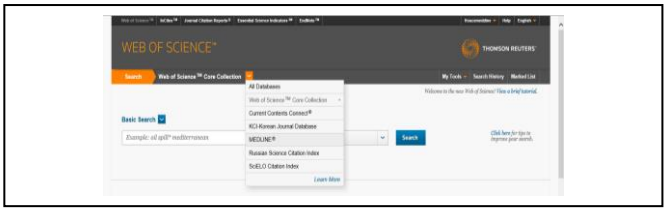


Fig. 3. Choosing Medline as the database of the Web of Science

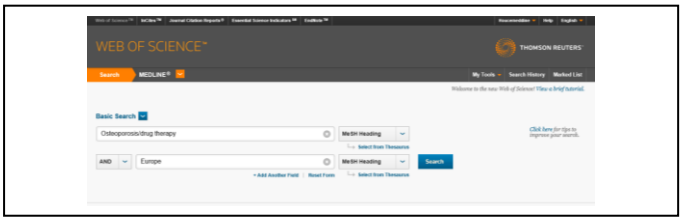


Fig. 4. Entering the Medline MeSH-based query to Web of Science

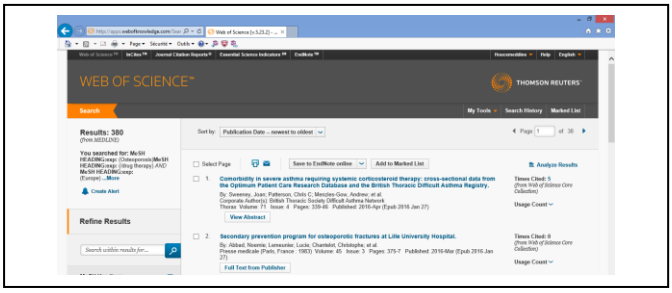


Fig. 5. Search results

If the search results exceed 100 papers, I choose the 100 papers that are mostly cited in the last year (2016 for a literature review written in 2017) from the initial set of search results. In fact, it seems that the yearly most cited papers about a given topic are known to the ones having the best quality [29, 30].

After I made the list of the best 100 papers and as the Medline database does not provide a list of the used references for each paper (Required for the selection of the main papers to be included in the given literature review using citation analysis), I will find them using Web of Science Core Collection as a database and add them to the Marked list of Web of Science website. If a paper does not exist in Web of Science Core Collection, it is substituted by the paper ranked just after it per its number of received citations in the last year.

After I add the list of considered papers to the Marked List, I click on the "Marked List" Button and retrieve the details of the papers (References, Authors, Titles…) as a plain text as shown in Fig. 6. This plain text will be used later in eliminating useless references using citation analysis.

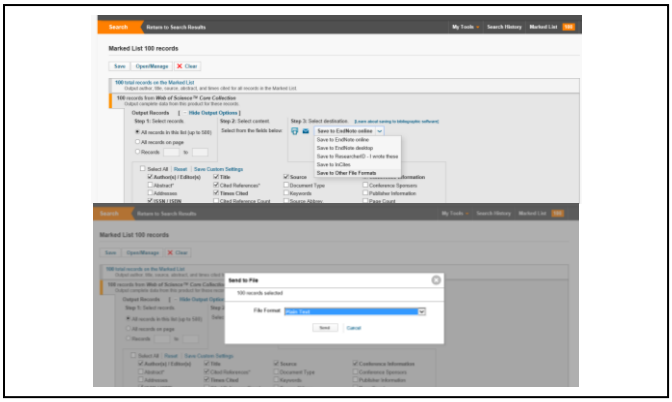


Fig. 6. Retrieving the details of the Marked List of Web of Science as a plain text

B. *Selection of the primary references*

As explained in [17], only the papers with high weight in the co-citation network or in the direct citation network of the topic are the primary references that should be used for the creation of most of the details of a given literature review.

In this workflow, I apply this method to choose the primary references to be considered for a given literature review. Effectively, I build the direct citation network and the co-citation network of the topic using VOSviewer, a software that was created by Van Eck and Waltman and that use a Web of Science plain text to generate citation networks (of authors, institutions, journals, or journals), co-citation networks (of authors, institutions, journals, or journals) or co-word networks (for abstracts, keywords, or titles of the papers) [31].

Then, I retrieve the list of primary references (papers with high weight in one of the two network). This is not difficult as shown in the sample co-citation network (Fig. 7) and direct citation network (Fig. 8).

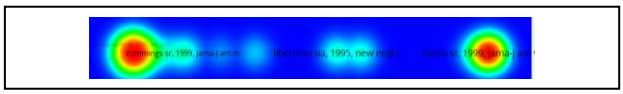


Fig. 7. Co-citation network (Density representation)

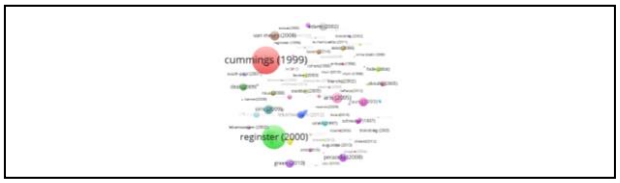


Fig. 8. Direct Citation network (Node representation)

C. *Selection of the secondary references*

As shown in [32], the automated method with high precision (F_1-test=0.793) that can be used to retrieve secondary references that explain the background or the findings of the primary considered references for a given literature review is snowballing technique. Snowballing technique consists of retrieving important papers about the topic having a direct or indirect citation link with the primary references [32].

Consequently, in this workflow, I build the direct citation network of the initially considered papers about the topic using CitNet Explorer [33]. Then, I consider the papers that are included in the independent components of the direct citation network involving the primary references and that have a significant weight within the direct citation network as the secondary references of the given literature review.

Using this method, the secondary references are not difficult to retrieve as shown in Fig. 9.

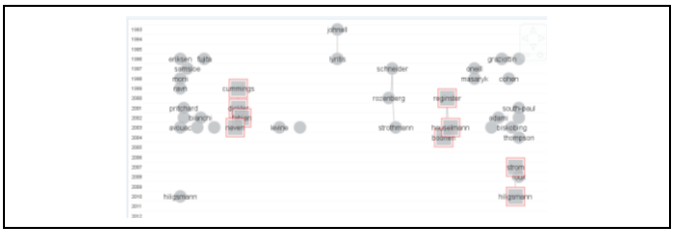


Fig. 9. Historiographic Direct citation network (Representation of the direct citation network in function of the years of publication of the papers)

III. Case Study: Drug Therapy of Osteoporosis in Europe

To assess the efficiency of this algorithm, I applied it to select references for a review about "drug therapy of osteoporosis in Europe".

A. *Speed of the method*

I used for that a 65 Megabits/s Internet connection and an ASUS computer with a 3.3 GHz Intel Atom processor and 1GB RAM. After entering the query to WoS, I obtained 389 search results in 0.8 seconds. Then, I ranked the search results per their number of 2016 citations to select the best 100 papers in 15 min 31 s and I spent 5 h 30 min in finding the 100 selected papers in WoS Core Collection and adding them as WoS Core Collection entries to the Marked List of Web of Science. I spent much time in these two steps because they cannot be automatically done using Web of Science. These two steps can be automated later in Web of Science by adding the "Times cited last year – Highest to lowest" option to "Sort By" service and by adding the "Add to Marked List as WoS Core Collection entries" option to "Add to Marked List" service.

Luckily, these two steps are the only ones that caused matters as the following steps of the workflow are automated and can be done quickly. In fact, I automatically retrieved the titles, authors, sources and cited references for the 100 papers in plain text format in 5.2 s. I succeeded using VOSviewer to retrieve the co-citation network in 1 min 08 s. and the direct citation network in 48 s. After that, I obtained the historiographic direct citation network using CitNet Explorer in 43 s.

B. *Precision of the results*

This workflow has selected 5 primary references (4 research articles and 1 review) and 11 secondary references (6 research articles and 5 reviews) to be cited in the biomedical literature review about "drug therapy of osteoporosis in Europe". When I analyzed the abstracts, the citations and full texts of these 16 papers using NLP techniques explained in works like [8] and [10], and [11], I found that all these references are unsurprisingly within the scope of the work (A paper cannot have a significant weight in the direct citation network of a topic outside its scope

[18]). This means that this method is precise and can be used alone to select references for a biomedical literature review without any risk of error (Using NLP analysis techniques to verify the results of this method is useless).

IV. Conclusion

In this paper, I propose a citation analysis-based complete workflow to drop irrelevant or useless papers from the initial set of considered references of a biomedical literature review using citation mapping technique.

Future directions of this research can be the substitution of the paid citation database used in this work (Web of Science Core Collection) by a free and wider citation database (Google Scholar), the amelioration of the speed of the workflow, or the creation of the software that automatically applies this workflow.

References

[1] G. Tsafnat, P. Glasziou, M. K. Choong, A. Dunn, F. Galgani and E. Coiera, "Systematic review automation technologies," *Systematic reviews*, vol. 3, no. 1, p. 1, 2014.

[2] L. Bornmann and H. D. Daniel, "What do citation counts measure? A review of studies on citing behavior," *Journal of Documentation*, vol. 64, no. 1, pp. 45-80, 2008.

[3] G. Tsafnat, A. Dunn, P. Glasziou and E. Coiera, "The automation of systematic reviews," *BMJ*, vol. 346, p. f139, 2013.

[4] B. C. Wallace, T. A. Trikalinos, J. Lau, C. Brodley and C. H. Schmid, "Semi-automated screening of biomedical citations for systematic reviews," *BMC bioinformatics*, vol. 11, no. 1, p. 1, 2010.

[5] J. Thomas and A. Harden, "Methods for the thematic synthesis of qualitative research in systematic reviews," *BMC medical research methodology*, vol. 8, no. 1, p. 1, 2008.

[6] P. Fontelo, F. Liu and M. Ackerman, "ask MEDLINE: a free-text, natural language query tool for MEDLINE/PubMed," *BMC Medical Informatics and Decision Making*, vol. 5, no. 1, p. 1, 2005.

[7] Evidence Partners, DistillerSR [Computer Program], Ottawa, Canada: Evidence Partners, 2011.

[8] D. Zhang, J. Lei and A. K. Robinson, "A hybrid approach for automating citation screening process," in *Abstracts of the 21st Cochrane Colloquium*, Québec City, 2013.

[9] A. Booth, "Searching for qualitative research for inclusion in systematic reviews: a structured methodological review," *Systematic reviews*, vol. 5, no. 1, p. 74, 2016.

[10] P. O'Blenis, U.S. Patent Application No. 11/271,805, Washington, DC: U.S. Patent and Trademark Office, 2005.

[11] C. J. Groot, T. Leeuwen, B. W. J. Mol and L. Waltman, "A longitudinal analysis of publications on maternal mortality," *Paediatric and perinatal epidemiology*, vol. 29, no. 6, pp. 481-489, 2015.

[12] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. A. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen and D. Moher, "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration," *Annals of internal medicine*, vol. 151, no. 4, pp. W-65, 2009.

[13] A. Y. Gasparyan, L. Ayvazyan, H. Blackmore and G. D. Kitas, "Writing a narrative biomedical review: considerations for authors, peer reviewers, and editors," *Rheumatology International*, vol. 31, no. 11, p. 1409, 2011.

[14] P. Cronin, F. Ryan and M. Coughlan, "Undertaking a literature review: a step-by-step approach," *British journal of nursing*, vol. 17, no. 1, p. 38, 2008.

[15] J. J. Randolph, "A guide to writing the dissertation literature review. Practical Assessment," *Research & Evaluation*, vol. 14, no. 13, pp. 1-13, 2009.

[16] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *Journal of the Association for Information Science and Technology*, vol. 24, no. 4, pp. 265-269, 1973.

[17] K. W. Boyack and R. Klavans, "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, no. 12, pp. 2389-2404, 2010.

[18] E. Garfield, "From the science of science to Scientometrics visualizing the history of science with HistCite software," *Journal of Informetrics*, vol. 3, no. 3, pp. 173-179, 2009.

[19] H. Keathley-Herring, E. Van Aken, F. Gonzalez-Aleu, F. Deschamps, G. Letens and P. C. Orlandini, "Assessing the maturity of a research area: bibliometric review and proposed framework," *Scientometrics*, vol. 109, no. 2, pp. 927-951, 2016.

[20] J. Shen, L. Yao, Y. Li, M. Clarke, L. Wang and D. Li, "Visualizing the history of evidence-based medicine: A bibliometric analysis," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 10, pp. 2157-2172, 2013.

[21] M. W. Bruner, K. Erickson, B. Wilson and J. Côté, "An appraisal of athlete development models through citation network analysis," *Psychology of Sport and Exercise*, vol. 11, no. 2, pp. 133-139, 2010.

[22] A. Kolchinsky, A. Abi-Haidar, J. Kaur, A. A. Hamed and L. M. Rocha, "Classification of protein-protein interaction full-text documents using text and citation network features," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 7, no. 3, pp. 400-411, 2010.

[23] S. C. Brailsford, P. R. Harper, B. Patel and M. Pitt, "An analysis of the academic literature on simulation and modelling in health care," *Journal of simulation*, vol. 3, no. 3, pp. 130-140, 2009.

[24] C. Chen, Z. Hu, S. Liu and H. Tseng, "Emerging trends in regenerative medicine: A scientometric analysis in CiteSpace," *Expert opinion on biological therapy*, vol. 12, no. 5, pp. 593-608, 2012.

[25] H. J. Lowe and G. O. Barnett, "Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches," *Jama*, vol. 271, no. 14, pp. 1103-1108, 1994.

[26] Thomson Reuters, Web of Science, 2016.

[27] L. Leydesdorff and T. Opthof, "Citation analysis with medical subject Headings (MeSH) using the Web of Knowledge: A new routine," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 5, pp. 1076-1080, 2013.

[28] D. Rotolo and L. Leydesdorff, "Matching Medline/PubMed data with Web of Science: A routine in R language," *Journal of the Association for Information Science and Technology*, vol. 66, no. 10, pp. 2155-2159, 2015.

[29] C. Lokker, K. A. McKibbon, R. J. McKinlay, N. L. Wilczynski and R. B. Haynes, "Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study," *Bmj*, vol. 336, no. 7645, p. 6, 2008.

[30] J. Manriquez, K. Cataldo and I. Harz, "Factors influencing citations to systematic reviews in skin diseases: a cross-sectional study through Web of Sciences and Scopus," *Anais brasileiros de dermatologia*, vol. 90, no. 5, pp. 646-652, 2015.

[31] N. J. Van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," *Scientometrics*, vol. 84, no. 2, pp. 523-538, 2010.

[32] M. K. Choong, F. Galgani, A. G. Dunn and G. Tsafnat, "Automatic evidence retrieval for systematic reviews," *Journal of medical Internet research*, vol. 16, no. 10, p. e223, 2014.

[33] N. J. van Eck and L. Waltman, "CitNetExplorer: A new software tool for analyzing and visualizing citation networks," *Journal of Informetrics*, vol. 8, no. 4, pp. 802-823, 2014.