DOI: http://dx.doi.org/10.15446/ing.investig.v40n1.77718

SAP Algorithm for Citation Analysis: An improvement to Tree of Science

Algoritmo SAP para análisis de citaciones: una mejora para Tree of Science

Daniel-Stiven Valencia-Hernández[1], Sebastián Robledo[2], Ricardo Pinilla[3], Néstor Darío Duque-Méndez[4], and Gerard Olivar-Tost[5]

ABSTRACT

Tree of Science (ToS) is a web-based tool which uses the network structure of paper citation to identify relevant literature. ToS shows the information in the form of a tree, where the articles located in the roots are the classics, in the trunk are the structural publications, and leaves are the most current papers. It has been found that some results in the leaves can be separated from the tree. Therefore, an algorithm (SAP) is proposed, in order to improve results in the leaves. Two improvements are presented: articles located in the leaves are from the last five years, and they are connected to root and trunk articles through their citations. This improvement facilitates construction of current literature for researchers.

Keywords: Tree of science, SAP, algorithm, citation analysis

RESUMEN

Tree of Science (ToS) es una herramienta web que usa la estructura de la red de citaciones para identificar literatura relevante. ToS muestra la información en forma de árbol, donde los artículos localizados en las raíces son los clásicos, en el tronco están las publicaciones estructurales y las hojas son los artículos más recientes. Se ha encontrado que algunos resultados de las hojas pueden ser separados del tema del árbol. Por lo tanto, se propone el algoritmo SAP para mejorar los resultados de las hojas. Se presentan dos mejoras: los artículos localizados en las hojas son de los últimos 5 años, y también, estos están conectados a la raíz y al tronco a través de sus citaciones. Esta mejora facilita la construcción de literatura actual a los investigadores.

Palabras clave: Tree of Science, SAP, algoritmo, análisis de citaciones

Received: February 7th, 2019
Accepted: January 30th, 2020

Introduction

Tree of Science (ToS) is a web-based tool that uses graph algorithms to optimize the search and selection of published papers. ToS was created at Universidad Nacional de Colombia (Robledo et al., Osorio-Zuluaga, and López-Espinosa, 2014), and the algorithm is explained elsewhere (Zuluaga et al., 2016). ToS is a specialized tool for researchers interested in tracking the way in which a particular topic evolves over time. Firstly, users must download Web of Science (WoS) query results. Then, they upload the file to ToS (tos.manizales.unal.edu.co). With this data, ToS shows the results in the form of a tree: root, trunk, and leaves. Papers in the roots are the classics, while those in the trunk are considered structural publications, and current papers are the leaves. In addition, ToS uses scientometric techniques to recommend relevant literature.

Scientometrics refer to the study of science, technology, and innovation from a quantitative perspective. Moreover, it focuses on the measurement of the impact of articles, journals, and institutions, along with the mapping of scientific areas (Leydesdorff, 2013). Examples include citation analysis (Koseoğlu, Sehitoglu, and Craft, 2015), co-author analysis (Ioannidis, 2015), and the impact of institutions (Singh, Uddin, and Pinto, 2015). Thus, the importance of scientometrics is based on the possibility of identifying high impact articles and main researchers, and on recognizing emerging areas of knowledge (Hood and Wilson, 2001).

[1] Systems Information Administrator, Universidad Nacional de Colombia. Colombia. Affiliation: Software Engineer, Vendoo, United States. Researcher in Core of Science, Colombia.E-mail: dsvalenciah@unal.edu.co

[2] Industrial Engineer, Universidad Nacional de Colombia, Colombia. M.B.A. Universidad Nacional de Colombia, Colombia. Ph.D. in Engineering, Universidad Nacional de Colombia, Colombia. Affiliation: Research-Professor, Universidad Católica Luis Amigó, Manizales. Colombia. ECOSOL Research Group. Director of Core of Science, Colombia. E-mail: sebastian.robledogi@amigo.edu.co

[3] Mathematician. Universidad Nacional de Colombia, Colombia. MSc. Applied Mathematics. Universidad Nacional de Colombia, Colombia. Affiliation: Professor, Universidad Nacional de Colombia, Colombia. E-mail: rpinillae@unal.edu.co

[4] Mechanical engineer. Universidad Nacional de Colombia, Colombia. MSc. Informatics. Ph.D. engineer. Universidad Nacional de Colombia, Colombia. GAIA Research Group. Affiliation: Professor, Universidad Nacional de Colombia, Colombia. E-mail: ndduqueme@unal.edu.co

[5] Mathematician. Ph.D. Applied Mathematics. Affiliation: Universidad de Aysén, Chile & Universidad Nacional de Colombia, Colombia & Arizona State University, USA. E-mail: gerard.olivar@uaysen.cl

How to cite: Valencia-Hernández, D. S., Robledo, S., Pinilla, R., Duque-Méndez. N. D., and Olivar-Tost, G. (2020). SAP Algorithm for Citation Analysis: An improvement to Tree of Science. *Ingeniería e Investigación*, *40*(1), 45-49. 10.15446/ing.investig.v40n1.77718

Scientometrics emerged in the 1930s with the analysis of the distribution frequency of productivity between chemistry and physics by Alfred J. Lotka (1926). After analyzing a number of publications, he concluded that the proportion of researchers making small contributions was 60%. Later, Derek J. de Solla Price (1963), known as the father of scientometrics, formulated Price's Law, which explains that 25% of scientific authors are responsible for 75% of published articles (preferential attraction model). Finally, another important initial contribution in this field was the h-index (Garfield, 1972; Hirsch, 2005), which measures the impact of papers and is well known in the scientific community nowadays. Consequently, these results showed patterns in the scientific world, which can be identified by mathematical and statistical analysis.

Currently, thanks to advances in technology, such as the Internet, it is possible to apply and develop sophisticated scientometric techniques in different fields. For example, a study in nanotechnology and nanoscience shows metrics such as the annual growth rate, authorship patterns, and an index of collaboration (Karpagam, Gopalakrishnan, Natarajan, and Ramesh Babu, 2011). Another investigation in bioenergy from biomass explains the exponential growth and changes in this field (Konur, 2012). Therefore, scientometrics have been a useful tool in recent years to identify emerging areas of science.

Although scientometrics have evolved in the last few years, one of the main challenges is to find accurate methods for the characterization of a scientific area (Koseoğlu, Sehitoglu, and Craft, 2015). For this reason, various researchers have proposed other indexes to determine the impact of publications. These include the CDS-index (Vinkler, 2011), multivariate analysis techniques, time series (Leydesdorff, 2013), and modeling techniques (Mutschke and Mayr, 2014). However, co-citation analysis has become a well-established topic in scientometrics to identify "sleeping beauty publications" (Fang, 2019 p. 307). Examples of applications of this scientometric techniques are found in reviews about obesity (Landinez, Robledo, and Montoya 2019), Corporate Social Responsibility (Duque and Cervantes-Cervantes, 2019), and in agriculture (Robledo-Buriticá, Aguirre-Alfonso, and Castaño-Zapata, 2019).

During the last years, some graph algorithms have been implemented in co-citation analysis to select relevant literature. For instance, HITS algorithm (Kleinberg, 1999) was applied to reduce ranking bias (Jiang et al. 2016) and Google's PageRank algorithm, to find the most prestigious papers (Chen et al. 2007). Nevertheless, much uncertainty still exists about tracking global knowledge using co-citation analysis (Parolo, Kujala, Kaski, and Kivela, 2019). Hence, this study seeks to improve the ToS algorithm to streamline the research process on a specific topic, in order to fulfill the need for non-conventional literature review techniques (Alulema and Largo, 2019) that other studies have proposed (Sepúlveda and Cravero, 2015).

This paper is structured as follows: First, a few basic definitions about graph theory are presented. Secondly, the methodology is described, detailing the algorithm step by step. Next, the SAP algorithm is applied to create a graph of citation analysis about Word-of-Mouth Marketing, in order to compare it with the current ToS results. Finally, conclusions are addressed, and limitations and implications are discussed.

Some basic definitions

Some basic definitions about graph theory are explained below, according to Johnsonbaugh (1999):

Definition 1 (undirected graph): A graph (or undirected graph) consists of a set of vertices V and a set of edges E, arranged in such a way that each edge $e \in E$ is associated with an unordered pair of vertices. If there is a unique edge e associated with the vertices v and w, it is written as follows: $e = (v, w)$ or $e = (w, v)$. In this context, (v, w) denotes an edge between v and w in an undirected graph and not an ordered pair.

Definition 2 (directed graph): A directed graph (or digraph) G consists of a set of vertices V and a set of edges E, arranged in such a way that each edge $e \in E$ is associated with an ordered pair of vertices. If there is a unique edge e associated with the ordered pair (v, w) of vertices, it is written as follows: $e = (v, w)$, which denotes an edge from v to w, where v is the initial vertex and w is the terminal vertex of the edge e.

Definition 3 (indegree and outdegree of vertex): Let v be a vertex of a directed graph G. The degree of entry of v, denoted by indegree (v), is the number of edges in G with terminal vertex v. The degree of output of v, denoted by outdegree (v), is the number of edges in G whose initial vertex is v.

Definition 4 (subgraph): Let $G = (V, E)$ a graph. $G' = (V', E')$ is a subgraph of G if:

1. V′ ⊆ V and E′ ⊆ E.
2. For each edge $e' \in E'$, if e' is incident on v' and w', then v', $w' \in V'$.

Definition 5 (connected graph): A graph G is connected if there is a walk between every pair of distinct vertices in the graph.

Definition 6 (connected component): A connected component of a graph G is a connected subgraph S of G such that no other connected subgraph of G contains S.

Data

In order to test the algorithm, we used data from Web of Science (WoS). This dataset contains information about articles published by journals from different areas of knowledge. From it, we can extract the citation relationships between papers, authors, publication dates, journals, volume, page, and the Digital Object Identifier (DOI). Similarly, we can create a citation graph with the papers and their references (Zuluaga et al., 2016).

SAP Algorithm

The SAP algorithm was implemented in Python with the graph package igraph. The operation of SAP is explained below.

Description

1. The SAP algorithm consists in six steps: From a subset of papers V, which is obtained from WoS, a directed graph $G = (V, E)$, with all the papers and references is generated, where each directed edge (i, j) of E is a citation from paper p_i to p_j.

2. Graph G is filtered:

   2.1 The largest connected component is obtained.

   2.2 Loops are eliminated from the graph obtained in (2.1).

   2.3 Duplicated edges are removed from the graph obtained in (2.2).

   2.4 Vertices with indegree 1 and outdegree 0 are eliminated, along with their edges, from the graph obtained in (2.3). This graph is noted by $G' = (V', E')$.

   Igraph Description:

   2.1 Graph.clusters(), which shows the different components of the graph, and the giant() function are used to select the largest component.

   2.2 and 2.3. Graph.simplify() is used to remove repeated loops and edges.

   2.3 Graph.vs.select() is used to select the vertices that do not have indegree 1 and outdegree 0.

3. Root classification:

   3.1 Vertices with outdegree 0 are selected from V'.

   3.2 V_{root} is defined as the set of all vertices obtained in (3.1).

   3.3 If r is a root, its SAP is defined as its indegree

   Igraph Description:

   3.1 Graph.vs.select() is used to choose the vertices with outdegree 0.

   3.2 Indegree() is used to determine the input degree

4. Leaves classification:

   4.1 Vertices with indegree 0 are selected from V'.

   4.2 $V_{extended\ leaf}$ is defined as the set of all vertices obtained in (4.1).

   4.3 Vertices whose age (time since publication) is not less than the newest vertex age less 5 of $V_{extended\ leaf}$, are selected from the vertices obtained in (4.1).

   4.4 V_{leaf} is defined as the set of all vertices obtained in (4.3).

   4.5 If v belongs to $V_{extended\_leaf}$, its SAP is defined as the number of pahts that exist between v and the roots.

   Igraph Description:

   4.1 Graph.vs.select() is used to choose the vertices with indegree 0.

   4.5 Graph.shortest_paths_dijstra() is used to identify paths between the vertices of $(V_{leaf} \cup V_{extended\ leaf})$ and the roots.

5. Trunk classification

   5.1 Vertices of V_{root} are selected.

   5.2 Vertices of (5.1) are sorted in descending order according to their SAP value.

   5.3 $V_{root\ selected}$ is defined as the first 10 vertices obtained in (5.2).

   5.4 Vertices of V_{leaf} are selected.

   5.5 Vertices of (5.4) are sorted in descending order according to their SAP value.

   5.6 $V_{leaf\ selected}$ is defined as the first 60 vertices obtained in (5.5).

   5.7 All the vertices that belongs to at least one path between $V_{root\ selected}$ and $V_{Leaf\ selected}$, are selected.

   5.8 V_{trunk} is defined as the vertices obtained in (5.7).

   5.9 If t is a trunk, its SAP is defined as the sum of the SAPs of the vertices that belong to $(V_{root\ selected}) \cup V_{leaf\ selected}$, and are connected with t by one or more paths.

   5.10 Vertices whose age (time since publication) is not older than the newest vertex of V_{trunk}, are selected from the vertices obtained in (5.8).

   5.11 $V_{potential\ leaf}$ is defined as the set of all the vertices obtained in (5.10).

   Igraph Description

   5.1 And 5.4 Graph.vs.select() is used to choose the vertices.

   5.2 Graph.get_all_simple_paths() is used to select the vertices between $V_{root\ selected}$ and $V_{leaf\ selected}$ to save trunk vertices and compute its SAP.

6. Tree construction: Subgraph $G = (V, E)$ of $G' = (V', E')$, where $V = (V_{root} \cup V_{leaf} \cup V_{trunk})$, and E is considered a subset of the edges of E' which only affects the vertices of V is called "Tree of Science" (Robledo et al. 2014).

Application

We compared the results from the ToS with the SAP algorithm to illustrate its operation. The similarities and differences between the two procedures are presented here.

The first step is to define the research topic, in order to obtain the data from WoS. In this case, Word-of-Mouth Marketing (WOMM) is used as the search equation for the time period from January 2001 to August 22, 2017.

Title = (marketing) AND Topic = (Word of Mouth) Indexes: SCI-EXPANDED, SSCI, A&HCI

Exactly 317 papers were extracted with these references. With this data, both algorithms were applied in order to identify the SAP improvements. Table 1 shows the difference between them. The results in the roots and trunk of both algorithms are similar: 90% and 70% respectively. However, ToS performs better in terms of number of citations: 2,819 results in the roots and 1,752 in the trunk. Despite this, SAP has an outstanding performance with more than three times the citation results from ToS.

Table 1. Differences between ToS and SAP algorithm

		Root	Trunk	Leaves
Similarities		90%	70%	23%
Differences in citations	ToS	2 819	1 752	741
	SAP	666	1 001	2 442

Source: Authors

Another important result of the SAP is the age of the papers on the leaves. According to Robledo et al. (2014), leaves are current papers, and a quality indicator of an investigation is the number of recent references. Moreover, Price (1976) suggests that at least 50% of the references should be from the past five years. The SAP meets this requirement by selecting papers from the past five years for the leaves (Figure 1).

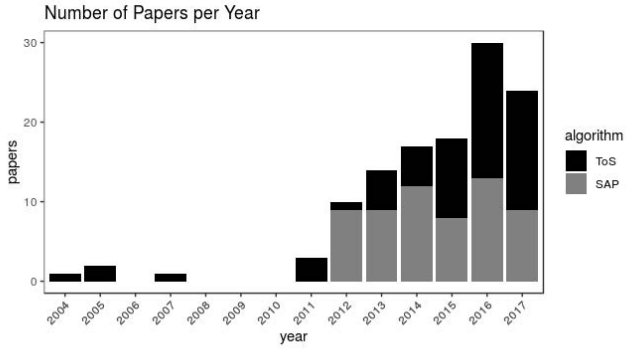

Figure 1. The number of paper per year per algorithm.
Source: Authors

Conclusions

ToS is a scientometric tool which performs the citation analysis of a graph and shows the results in a form of a tree: root, trunk, and leaves. Most of the results presented by ToS are relevant and important (Robledo et al. 2014). However, there is a lack of precision in the leaves; sometimes publications are not connected to the roots and trunk, and additionally, the leaves are occasionally not current literature. Thus, the goal of this study is to propose a new algorithm called SAP, which improves the results in the leaves.

Results show that SAP is more accurate in this field. It presents the most important current literature. However, this study is limited, and so it must be further expanded, for example, to the evaluation both of different research topics and indicators, in order to understand the pros and cons of the new algorithm.

Acknowledgements

We thank the "Instituto Colombiano para Ciencia y Tecnología" (Colciencias) and its program "Jóvenes investigadores e innovadores". Also, this investigation was partially supported by Universidad Nacional de Colombia under project No. 32059 (Code Hermes), entitled "Consolidación de las líneas de investigación del Grupo de Investigación en Ambientes Inteligentes Adaptativos GAIA" within the "Convocatoria interna de investigación de la Facultad de Administración 2015, para la formulación y ejecución de proyectos de consolidación y/o fortalecimiento de los grupos de investigación. Modalidad 1: Formulación y ejecución de proyectos de consolidación". We also would like to thank the Universidad Nacional -Sede Manizales library for its diffusion of ToS, and their support in this project. We thank Universidad Católica Luis Amigó and the research group ECOSOL for the time given to finish this paper. Gerard Olivar-Tost thanks Colciencias for supporting the publication of this paper through the project "Modelado y simulación del Metabolismo Urbano de Bogotá D.C. Código 111974558276". We thank the anonymous reviewers for their insightful suggestions to the manuscript and Oscar David Arbeláez Echeverri and Juan David Alzate Cardona for their support in the algorithm implementation. The repository of the code is in this link https://github.com/coreofscience/python-sap

References

Alulema, F. X. V. and Largo, F. L. (2019). Strategic portfolio of IT projects at universities: A systematic and non-conventional literature review. *Ingeniería e Investigación*, *39*(2). 10.15446/ing.investig.v39n2.72431

Chen, P., Xie, H., Maslov, S., and Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, *1*(1), 8-15. 10.1016/j.joi.2006.06.001

Duque, P. and Cervantes-Cervantes, L. S. (2019). Responsabilidad Social Universitaria: una revisión sistemática y análisis bibliométrico. *Estudios Gerenciales*, 35(153), 451-464. 10.18046/j.estger.2019.153.3389

Fang, H. (2019). A transition stage co-citation criterion for identifying the awakeners of sleeping beauty publications. *Scientometrics*, 121(1), 307-322. 10.1007/s11192-019-03195-9

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science, 178*(60), 471-479. 10.1126/science.178.4060.471

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569-72. 10.1073/pnas.0507655102

Hood, W. W. and Wilson, C. S. (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics, 52*(2), 291-314. 10.1023/A:1017919924342

Ioannidis, J. P. A. (2015). A generalized view of self-citation: Direct, co-author, collaborative, and coercive induced self-citation. *Journal of Psychosomatic Research, 78*(1), 7-11. 10.1016/j.jpsychores.2014.11.008

Jiang, X., Sun, X., Yang, Z., Zhuge, H., and Yao, J. (2016). Exploiting heterogeneous scientific literature networks to combat ranking bias: Evidence from the computational linguistics area. *Journal of the Association for Information Science and Technology*, *67*(7), 1679-1702. 10.1002/asi.23463

Johnsonbaugh R. (1999). *Discrete Mathematics*, New York: Macmillan Publishing Company.

Karpagam, R., Gopalakrishnan, S., Natarajan, M., and Ramesh Babu, B. (2011). Mapping of nanoscience and nanotechnology research in India: A scientometric analysis, 1990-2009. *Scientometrics, 89*(2), 501-522. 10.1007/s11192-011-0477-8

Kleinberg, J. M. (1999). Hubs, authorities, and communities. ACM computing surveys (CSUR), 31(4es), 5. 10.1145/345966.345982

Konur, O. (2012). The scientometric evaluation of the research on the production of bioenergy from biomass. *Biomass and Bioenergy, 47*, 504-515. 10.1016/j.biombioe.2012.09.047

Koseoğlu, M. A., Sehitoglu, Y., and Craft, J. (2015). Academic foundations of hospitality management research with an emerging country focus: A citation and co-citation analysis. *International Journal of Hospitality Management, 45*(0), 130-144. 10.1016/j.ijhm.2014.12.004

Landinez, D., Robledo, S., and Montoya, D. (2019). Executive Function performance in patients with obesity: a systematic review. *Psychologia*, 15(2), 31-50. 10.1007/s11192-012-0917-0

Leydesdorff, L. (2013). Statistics for the dynamic analysis of scientometric data: The evolution of the sciences in terms of trajectories and regimes. *Scientometrics*, *96*(3), 731-741. 10.1007/s11192-012-0917-0

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16*(2), 317-324.

Mutschke, P. and Mayr, P. (2014). Science models for search: a study on combining scholarly information retrieval and scientometrics. *Scientometrics, 102*(3), 2323-2345. 10.1007/s11192-014-1485-2

Parolo, P. D. B., Kujala, R., Kaski, K., and Kivela, M. (2019). Going beneath the shoulders of giants: tracking the cumulative knowledge spreading in a comprehensive citation network. *arXiv preprint arXiv:1908.11089*.

Price, D. J. de S. (1963). *Little science, big science and beyond*. Columbia University Press. New York. 10.7312/pric91844

Price, D. J. de S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American society for Information Science*, *27*(5), 292-306. 10.1002/asi.4630270505

Robledo-Buriticá J., Aguirre-Alfonso, C. A., and Castaño-Zapata, J. (2019). *Guía ilustrada de enfermedades en postcosecha de frutas y verduras y sus agentes causantes en Colombia*. Bogotá, Colombia: Academia Colombiana de Ciencias Exactas, Físicas y Naturales.

Robledo, S., Osorio, G. and López, C. (2014). Networking en pequeña empresa: una revisión bibliográfica utilizando la teoría de grafos. *Revista vínculos*, 11(2), 6-16. Retrieved from: https://revistas.udistrital.edu.co/ojs/index.php/vinculos/article/view/9664

Sepúlveda, S. and Cravero, A. (2015). Protocol adaptations to conduct systematic literature reviews in software engineering: A chronological study. *Ingeniería e Investigación*, *35*(3), 84-91. 10.15446/ing.investig.v35n3.46616

Singh, V. K., Uddin, A., and Pinto, D. (2015) Computer science research: the top 100 institutions in India and in the world. *Scientometrics* 104, 529-553. 10.1007/s11192-015-1612-8

Vinkler, P. (2011). Application of the distribution of citations among publications in scientometric evaluations. *Journal of the American Society for Information Science and Technology, 62*(10), 1963-1978. 10.1002/asi.21600

Zuluaga, M., Robledo, S., Osorio-Zuluaga, G. A., Yathe, L., Gonzalez, D., and Taborda, G. (2016). Metabolómica y Pesticidas: Revisión sistemática de literatura usando teoría de grafos para el análisis de referencias. *Nova, 13*(25), 121-138. 10.22490/24629448.1735