Topical Impact Analysis

A New Informetric Indicator for the Assessment of a Scientific Institution

Axel Honka, Lisa Orszullok, Isabelle Dorsch, Nils Frommelius

Heinrich Heine University
Universitätsstraße 1, Düsseldorf, Germany
{Axel.Honka, Lisa.Orszullok, Isabelle.Dorsch, Nils.Frommelius}@hhu.de

Abstract

This paper presents new specific values regarding the topical impact analysis. These values make comprehensible statements that provide concrete comparative output to describe the differences between an initial topical map and an impact map. The purpose of this study is to evaluate the topical scope any institution has with its output. The Know-Center, an institute for knowledge technology in Graz, acts as a case study. To collect the citations, we used the reference search at Scopus. All publications of the Know-Center from 1st January 2003 until 31th December 2012 have been covered. According to the study, there is a high rate of new topics in the impact whereat durable topics have a higher occurrence comparing to the Know-Center's output.

Keywords: Topical impact analysis, Informetric indicator, Scientific institution, Citation, Topic

In: F. Pehar/C. Schlögl/C. Wolff (Eds.). Re:inventing Information Science in the Networked Society. Proceedings of the 14th International Symposium on Information Science (ISI 2015), Zadar, Croatia, 19th–21st May 2015. Glückstadt: Verlag Werner Hülsbusch, pp. 410–421.

1 Introduction

Scientometric evaluation is a very important part of science. It has its roots back in 1963 in a book called "Little Science, Big Science" written by Derek de Solla Price who was later entitled as "the father of scientometrics" by Eugene Garfield and Robert Merton (de Solla Price, 1986). Van Raan (1997) defines scientometrics as "[…] quantitative studies of science and technology". These quantitative studies utilize different scientometric indicators that aid in making a specific research measureable, may it be for quantitative or qualitative purpose. Scientometric analyses offer the possibility of crediting influence by investigating the flow of information. The information flow is represented by citations und they uncover the link between scholarly work in form of articles (Garfield, 1964; Garfield, 1979). Most of the traditional bibliometric indicators focus on quantitative measurements based on publication and citation counts which is an approved method to evaluate e.g. authors, research groups, institutions. But scientometrics goes beyond this point and allows a deep insight into the impact of scholarly work and behaviour as well as development of research topics.

There are some studies on informetric and scientometric topic analysis (Mann, 2006). The empirical basis for topic analyses are either terms from the title (used, e.g.,by Milojević, Sugimoto, Yan, & Ding, 2011), terms from the full text of the publication, or terms from the document's metadata. Stock (1989; 1990a; 1990b) applied metadata derived by the text-word method (Stock & Stock, 2013: 735 ff.), but we can also use descriptors from a thesaurus, notations from a classification system, etc. To construct topical clusters, we have to choose appropriate similarity measures (as Jaccard-Sneath, Dice, or Cosine; see Heck, 2011), clustering algorithms (as single linkage, complete linkage or group average method; see Rasmussen, 1992; Stock & Stock, 2013: 777–779) and suitable threshold values. Additionally, we are in need of quantitative indicators to describe relations between different clusters (e.g., the cluster of the original literature and the cluster of the citing literature). Stock works with the reception-degree of terms (Stock, 1990b: 1299) and with the stability of clusters in the citing literature over time (Stock, 1989; Stock, 1990b: 1304).

1.1 What is new?

In this paper we introduce a new scientometric indicator regarding topical impact analysis along with several new specific values to describe the development and transformation of topical maps. A topical map is composed of the most frequent topics and its interconnectivity among themselves of any institution's research. Therefore topical impact maps focus on frequent topics of literature that have a reference on the primary documents of any institution.

The analysis of topical impact is a research area of informetrics that has a significant value for scientific institutions. Based on the analysis of topic networks we developed an indicator that allows scientific institutions to get an overview whereto their scholarly work reach and how it affects other researchers. It becomes clear in what specific topics the initial work reaches and how these themes are connected to each other.

The topical impact analysis is very flexible regarding its use. It can be applied to the entire output of a scientific institution, on city- or even country-level or for example to all publications on any author in order to measure his personal topical impact.

How can you measure the topical impact of institutions? To make a comprehensible statement about the impact, there need to be specific values that provide concrete comparative output measurements to document the development of the primary topic network to the cited impact map.

1.2 New specific values

The first value for investigating topical maps is the single-topic dispersion value (STDV). What is its purpose and how is it utilized? This value compares an identical topic of two topical maps, the initial one and the impact one, regarding the total topic occurrence. With the help of this value, you can get a precise outcome of the topic development. The occurance in the topical map of the publication's impact can either rise which indicates a certain importance concerning the up-to-dateness of that particular topic, it can remain the same or the occurrence can decrease. This value is calculated by dividing the occurrence of the impact's topic through the initial one. A value of 1 expresses no change in terms of the occurrence. Values above 1 possibly denote a topic of interest in near future whereas values below 1 imply the opposite.

In the following, we introduce the second value, the new topical accrual value (NTAV). It considers topics that are not given in the original research cluster but appears in the citatory publications and are part of the impact network. Therefore it bears in mind that research results can be interdisciplinarily perceived. The formula of the new topical accrual value is as per particulars given below:

$$NTAV = (n - g) / n,$$

where n is the total count of topics in the impact-sided topical map and g the amount of equal topics in both topical maps. What does this value reveal? It states the ratio of "new" topics in the topical map of the publication's impact where "new" topics are defined as those topics that were not part of the initial topical map. The following showcase example shows how this value works: In the initial topical map there are ten different topics. The topical map of the impact has five topics in common but consist of ten new topics which results in a total size of fifteen. On appliance of the previously explained value this would result in as the following calculation shows:

$$NTAV = (15 - 5) / 15 = 0.666$$

In percentage, that would be 66% new topics compared to the initial map with regard to the size of the impact's topical map.

The last value to be introduced is the topical durability value (TDV). As the name says it is being used to make a statement about the durability of all initial topics. To be exact it states the percentage of how many publication's topics occur in the topical map of the impact, therefore any institution may see if topics that they focus are being picked up by the citations. It is calculated the same way as the new topical accrual value apart from the variable n which is now the total count of the initial topical map.

In this paper the topical impact analysis is illustrated by the Know-Center, an institute for knowledge technology in Graz, Austria. However each of the values can be applied to any institution and the impact of it.

2 Methods

The foundation of this research is formed by the study of Dorsch & Frommelius (2015). They collected all publications of three different institutions of information science in Graz, Austria: the evolaris next level GmbH, the

Know-Center and the Institute of Information Science and Information Systems at the Karl-Franzens-University. The study covered all publications of the scientists from 1st January 2003 until 31st December 2012.

2.1 Primary database and collection of citations

In order to identify the citations, we used the collection of all publications from the Know-Center. This was done with help of the "Cited-Reference-Search" in the Scopus-database. Though, Scopus does not include all publications. Every publication has been investigated for its citation that had an entry in Scopus at their state of research. Recently added publications were not included. Our observation period started at 5th June 2014 and ended at 1st September 2014.

There are some noteworthy features that have to be clarified in order to document the exact procedure of the research. In some cases the reference search delivered identical citations of publications from the Know-Center. All duplicates of any publication have been deleted from the collection of citations. It is important to clarify how we defined duplicates. If a publication has several citing publications with the same title it does not make them automatically a duplicate. They are invariably declared as duplicates if all citation information like author(s), year of publication, source title and document type are identical. If any of these information differ from each other regarding two quoting publications both are relevant for the analysis. Another feature has been citations whose title was not in English. Normally there is an equivalent translation generated by Scopus but some titles were disregarded. These citations have been manually translated into British English. The topical analysis and the topical maps have been created according to the rules of Dorsch & Frommelius (2015). The last methodological aspect is the feature of the adjusted and unadjusted collection of the citations. In the course of research, it turned out that there may be a problem with citations that appear more than once with regard to the topical map. Talking about duplicates before, we were located on the level of citations of any specific publication and its duplicates. Now we step up a level and talk about multiple appearing citations on the level of the citations of the entire amount of publications. Why did we differ between these two features? It was striking that there were citations that appeared multiple times since they had references to several publications of the Know-Center. The following exaggerated, fictitious example shows why there has to be distinguished between two different

collections of citations, the adjusted and the unadjusted one: There are 100 initial publications and in total they reach a citation count of 200. Each publication has a specific citation C and any other one. Two different terms T1 and T2 only appear together in citation C. In the topical map that would be a topical similarity of 1. This phenomenon should be critically questioned. Is it correct to say that term T1 and T2 have a very close connection to each other? Just in a limited way since the terms only appear in a specific citation. This is why there are two different types of databases containing the amount of citations and therefore two different topical maps which will be shown in the following section of this paper. Thus, solely the adjusted version was considered in our calculations. To construct the topical clusters, we applied the Jaccard-Sneath coefficient for similarity calculation and single linkage for clustering.

3 Results

This paragraph illustrates the results of the topical impact analysis based on the publications and its citations of the Know-Center.

Figure 1 and 2 illustrate the topical impact map of the Know-Center in two different versions, the unadjusted and the adjusted one. At first sight both diagrams share similarities but basically they differ from each other. The threshold for the topical similarity has been differently determined, the unadjusted version has a threshold of 0.1 and the adjusted one of 0.05 since there is only one topical similarity above the threshold of 0.1. Lowering the threshold of the adjusted version would double the amount of topical similarities which would make the topical map unclear.

Figure 1 shows 18 topics subdivided into six single-linkage parts. The topic with the highest occurrence is “work integrated learning” with a total count of 25. The two phrases “small enterprise” and “medium enterprise” have the highest topical similarity (1). This is because the two terms are often used as a phrase (“small and medium enterprise” or SME). According to the rules this phrase had to be split up into two separate ones. The second highest topical similarity (0.368) is given by the terms “retrieval” and “semantic web.”

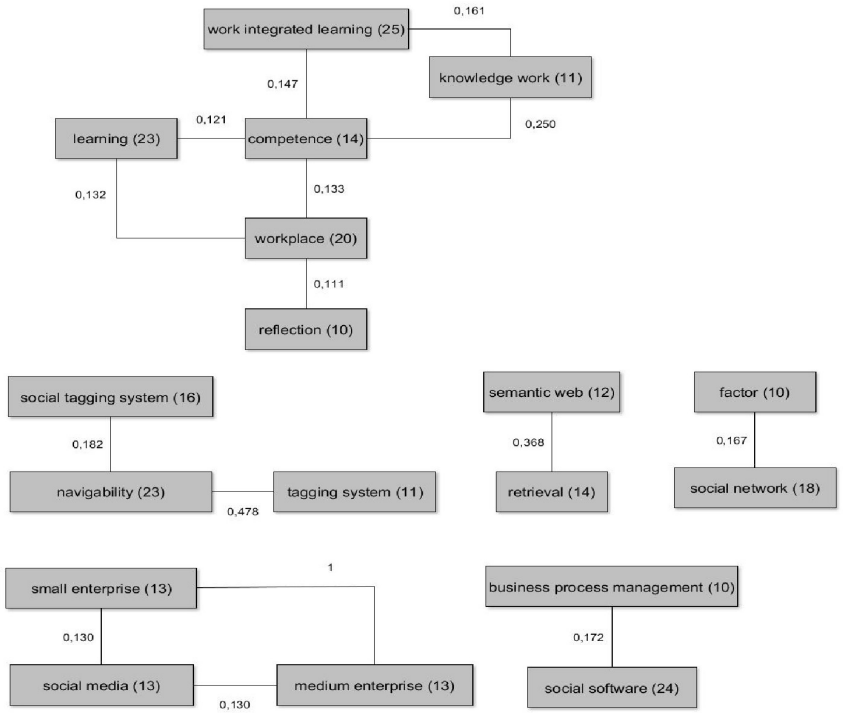

Figure 1. Unadjusted version of the topical impact map of the Know-Center

Figure 2 shows 20 topics which are subdivided into five single-linkage parts. Similar to figure 1 “small enterprise” and “medium enterprise” share the highest topical similarity. None of the others reach a value above 0.1. Terms and phrases with the highest occurrence (17) are “social software”, “learning” and “folksonomy”. Comparing the terms and phrases of both figures, they only have 6 terms or phrases in common (“learning”, “work-place”, “small enterprise”, “medium enterprise”, “social media” and “social software”). This emphasizes the necessity of differentiating between the adjusted and unadjusted version as explained in the method part. The top phrase “work integrated learning” with a count of 25 in the initial map did not even show up in the impact’s map since there were 17 duplicates on the level of all publications that had to be removed resulting in a count of 8 and therefore the phrase is not any longer relevant for the impact.

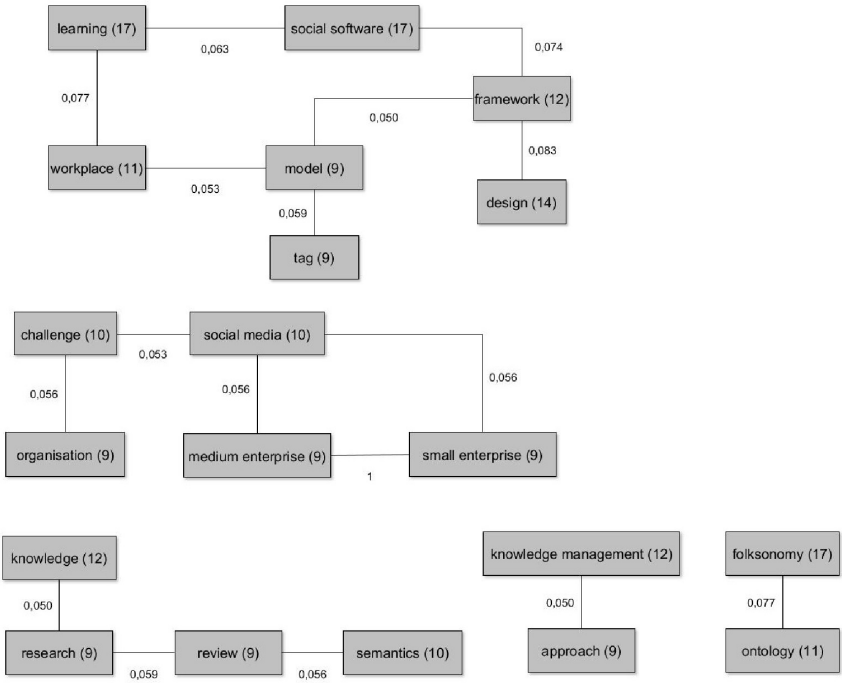

Figure 2. Adjusted version of the topical impact map of the Know-Center

In the following, all introduced specific values are exemplified on the publication's topical map (fig. 3) and its topical impact maps (fig. 1 and 2).

For the single-topic dispersion value we need to scan the initial topical map and the one of the impact for a term or phrase that appears in both maps. The phrase "social software" has in the initial topical map a value of 5, the same phrase has an amount of 17 in its impact. Dividing the impact's occurrence with the initial one we get a value of 3.4 which is more than three times the amount of the initial occurrence. An example for a constant value of 1 is the term knowledge. There is no negative example.

The NTAV gives a ratio of new topics to old ones in the impact's topical map. The amount of topics appearing in both maps is 6 ("workplace," "learning," "social software," "small company/enterprise," "medium company/enterprise" and "knowledge"). The size of the impact's map is 20. This leads to a value of 0.7 which makes out 70% new topics in the impact.

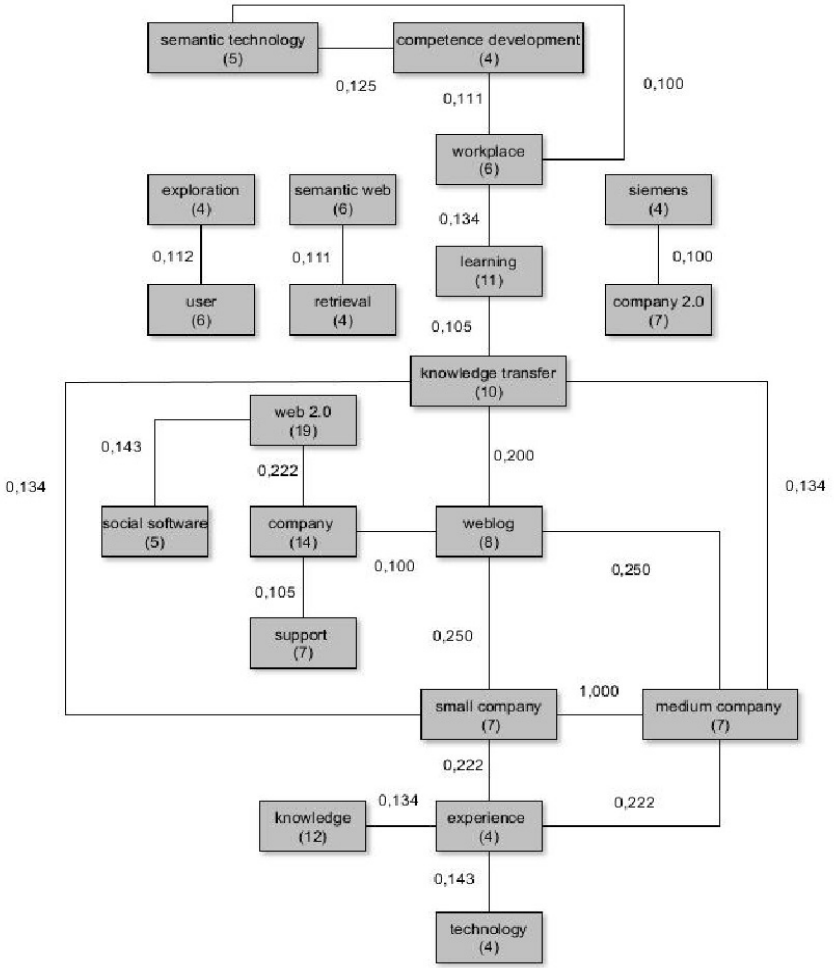

Figure 3. Topical map of the publications of the Know-Center

At last we have a look at the topic durability value. We have already mentioned that the number of topics that appear in both maps is 6. The size of the initial map is 21. Inserting these values in the formula which was presented in the introduction we get a result of 0.714. Therefore 71.4% of the initial topics do not appear in the impact map.

4 Discussion

In this paper we have shown the topical impact analysis combined with quantifiable topic network indicators. For this we have introduced several new specific values that describe the development of the impact-sided topical map regarding the initial map.

These values give valuable information about the reception of a certain output. This may be very interesting for any institution to see what kind of impact their research they are specified on has. This is indicated by the topic-durability value which illustrates the percentage of how many topics are being used for further research and how many are left behind. Each institution must decide on its own what percentage of topic loss is acceptable for them. Hence the limit of the percentage is flowing and has to be set individually.

The advantage of the single topic dispersion value is a clear statement about the development of any specific topic. Will this certain topic play a role in future research? Is there any nameable impact of this topic? Values above 1 indicate important topics with ongoing research.

The new topical accrual value is an indication in which topics the initial output operates. This is important for the institutions because they do not have an overview about the impact of their publications. So they merely do not only see the topical scope of their work, they see in which constellation these topics are connected with each other in the topical map.

Having a look at the result of the Know-Center, you can say that topics which remain in the impact tend to have a higher occurrence than in the initial map. In general the durability of topics is rather low since 71.4% of initial topics are not relevant in the impact.

As pointed out in the introduction the topical impact analysis is a very flexible indicator. It can be applied along with the specific values to any institution's output.

For further research, a more extensive study may be useful in order to develop a general scale. On the basis of that result, one can state whether a value is rather good or bad. However, there are lots of open questions: Which influence has the used citation database (Web of Science and Scopus) on the content of the topical clusters? What role do the similarity coefficients play, what the clustering method, and what the threshold values? Is it really helpful to work only with the publications' title terms, or do we need further information from the full text or from the metadata?

In addition, it would be interesting to see how the topical maps would develop across several more generations. Will it be similar to the behavioral pattern of the Know-Center in which only a small amount of topics play a role in the impact but those topics that remain are of higher importance with regard to its occurrence?

References

de Solla Price, D. (1963). *Little Science, Big Science.* New York, NY: Columbia University Press.

de Solla Price, D. (1986). *Little Science, Big Science ... and Beyond.* New York, NY: Columbia University Press.

Dorsch, I., & Frommelius, N. (2015). A scientometric approach to determine and analyze productivity, impact and topics based upon personal publication lists. In: F. Pehar, C. Schlögl & C. Wolff (Eds.). *Re:inventing Information Science in the Networked Society. Proceedings of the 14th International Symposium on Information Science – ISI 2015* (pp. 578–580). Glückstadt: Verlag Werner Hülsbusch.

Garfield, E. (1964). Science citation index, a new dimension in indexing. *Science*, 144 (3619), 649–654.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics*, 1 (4), 359–375.

Heck, T. (2011). A comparison of different user-similarity measures as basis for research and scientific cooperation. In: *ISSOME '11, Information Science and Social Media – International Conference, Åbo/Turku, Finland, August 24–26, 2011.*

Mann, G. S. (2006). Bibliometric impact measures leveraging topic analysis. In: *JCDL '06. Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 65–74). IEEE.

Milojević, S., Sugimoto, C. R., Yan, E., & Ding, Y. (2011). The cognitive structure of library and information science. Analysis of article title words. *Journal of the American Society for Information Science and Technology,* 62 (10), 1933–1953.

Rasmussen, E. M. (1992). Clustering algorithms. In: W. B. Frakes & R. Baeza-Yates (Eds.). *Information Retrieval. Data Structures & Algorithms* (pp. 419–442). Englewood Cliffs, NJ: Prentice Hall.

Stock, W. G. (1989). Datenbank „Grazer Schule". Eine Spezialdatenbank im Bereich der Philosophie- und Psychologiegeschichte. *Zeitschrift für philosophische Forschung,* 43, 347–364.

Stock, W. G. (1990a). Themenanalytische informetrische Methoden. In: M. Stock & W. G. Stock (Eds.). *Psychologie und Philosophie der Grazer Schule* (pp. 7–31). Amsterdam, The Netherlands, Altlanta, GA: Rodopi.

Stock, W. G. (1990b). Psychologie und Philosophie der Grazer Schule. Ein informetrischer Überblick zu Werk und Wirkungsgeschichte von Meinong, Witasek, Benussi, Ameseder, Schwarz, Frankl und Veber In M. Stock & W. G. Stock (Eds.). *Psychologie und Philosophie der Grazer Schule* (pp. 1223–1445). Amsterdam, The Netherlands, Altlanta, GA: Rodopi.

Stock, W. G., & Stock, M. (2013). *Handbook of Information Science.* Berlin, Boston: De Gruyter Saur.

van Raan, A. F. (1997). Scientometrics: State-of-the-Art. *Scientometrics*, 38 (1), 205–218.