Developing and applying a metric for estimating trends in databases of articles: Its possible use in research and science policy

JONATHAN M. LEVITT [1]

1 School of Library, Archive, and Information Studies
University College London, Gower Street, London WC1E 6B, U.K.

Abstract

The research presented very briefly in this paper aims to develop and apply a metric for estimating trends in databases of articles. The motivation for estimating trends in these databases is that these trends are possibly of interest in research and science policy.

The objectives of this paper are to introduce the metric (called 'absolute term incidences'), and to compare this metric with the metric used by Braun and Schubert and Gu (called 'term frequencies'), in addition to briefly present some potential uses and limitations of absolute term incidences. This research is part of my Doctoral project on bibliometrics and policy.

1. Introduction

This research was motivated by Henry Small's Paradigms, citations, and maps of science: A personal history ([1]). Small recommends measuring change through using "a variety of tools and units of analysis, for example, scientific terminology, indexing terms, cited documents, cited authors, questionnaires, and surveys." Small's recommendation persuaded the author of this paper that metrics for estimating trends could provide results useful to decision makers in research and science policy.

This research relates to Braun and Schubert's A quantitative view on the coming of age of interdisciplinarity in the sciences 1980-1999 ([2]) and Gu's Global knowledge management research: A bibliometric analysis ([3]). The research described in this presentation seeks to build on the approach used by Braun and Schubert, and Gu with a view to forming a method for estimating trends in databases of articles.

1.1. Methods

The metric used by Braun and Schubert [2] and Gu [3], called 'term frequencies', is the reported number of matches of a search query. This paper uses term frequencies and another metric, called 'absolute term incidences'. The word 'absolute' is used as a qualifier to 'term incidence', as the Doctorial research also uses the metric called 'relative term incidences'. Both absolute and relative term incidences are calculated by normalising term frequencies, but the way in which this normalisation is implemented varies according to the type of term incidence. This paper confines itself to absolute term incidences. Absolute term incidences are calculated by dividing term frequencies by the number of items searched and multiplying by 100,000.

For all the data investigated in this paper both term frequencies and absolute term incidences are presented. One advantage of using absolute term incidences as a metric is that, unlike term frequencies, they adjust for changes in size of the repository. Another advantage is that because absolute term incidences adjust for the number of items searched, they can also be used to compare how trends vary from repository to repository. One advantage of using term frequencies as a metric is that the size of the term frequency indicates the importance to attach to any findings on absolute term incidences.

1.2. Results

Braun and Schubert used the annual values of the term frequency of 'interdisciplin* AND multidisciplin*' in the titles of articles on ISI's Science Citation Index to estimate the growth in the extent of disciplinarity in research. Gu used the annual values of the term frequency of 'knowledge management OR knowledge discovery OR knowledge sharing' in the titles of articles on ISI's Web of Science to estimate the growth in knowledge management research.

In the Doctoral project, Braun and Schubert's research is compared with Gu's research. One finding was that the growth of the absolute term incidence of 'interdisciplin* AND multidisciplin*' only grew by about a factor of two in two decades. We considered whether the growth of 'interdisciplin* AND multidisciplin*' provides a close estimate of the growth in the extent of disciplinarity. We concluded that a high percentage of articles that are interdisciplinary or multidiscipliary in nature might not contain these words in their titles.

This paper investigates the growth of 'object orient*' and 'nanotechn*' on EBSCO's Academic Search Premier, a Web based database of articles. One reason for choosing to investigate 'object oriented' and 'nanotechnology' is that it was envisaged that the usage of these terms would have grown dramatically over the past 15 years. The major advantage of investigating trends on Academic Search Premier is that this database is searchable by month and year, whereas databases of articles are generally only searchable by year.

The tables presented in this paper have been selected in order to introduce the metrics. The use of these metrics is examined much more fully in the Doctoral project.

The term frequencies and absolute term incidences of 'object orient*' in the title of articles on Academic Search Premier were calculated for the period 1977-87, and for the years from 1988, and are presented in Table 1. The reason for starting in 1977 is that the first article on that database with 'object orient*' in the title was published in 1977.

Table 1: Term frequencies and absolute term incidences of 'object orient*' in the title (Academic Search Premier).

year	term frequency	absolute term incidence
2005	57	4.96
2004	73	5.94
2003	87	7.22
2002	60	6.36
2001	38	5.57
2000	46	6.96
1999	35	5.53
1998	48	7.68
1997	57	9.11
1996	59	9.96
1995	56	10.08
1994	47	10.93
1993	36	11.56
1992	17	7.61
1991	10	5.06
1990	12	7.01
1989	12	10.37
1988	9	11.24
1977-87	8	1.55

Although the annual term frequencies were substantially higher for the period 1993-2005 than for 1988-91, these differences were not present in the term incidences (The lowest annual term frequency for 1993-2005 was 35, whereas the highest annual term frequency for 1988-91 was 12; the annual absolute term incidence for 1993-2005 was between 4.96 and 11.56 whereas for 1988-91 it was between 5.06 and 11.24). In essence, the growth in frequency of 'object orient*' can be accounted for

by the growth in size of the repository. This illustrates the rationale for using normalised term frequencies to measure growth.

The term frequencies and absolute term incidences of 'nanotechn*' in the title of articles on Academic Search Premier were calculated for the period 1986-90, and for the years from 1991, and are presented in Table 2. The reason for starting in 1986 is that the first article on that database with 'nanotechn*' in the title was published in 1986.

Table 2: Term frequencies and absolute term incidences of 'nanotechn*' in the title (Academic Search Premier).

year	term frequency	absolute term incidence
2500	227	19.77
2004	203	16.52
2003	147	12.19
2002	70	7.42
2001	22	3.23
2000	24	3.63
1999	11	1.74
1998	11	1.76
1997	4	0.64
1996	4	0.67
1995	1	0.18
1994	5	1.16
1993	1	0.32
1992	4	1.79
1991	9	4.55
1986-90	4	0.80

The annual absolute term incidences were substantially higher for the period 2003-05 than for 1991-2001 (The annual absolute term incidence for 2003-05 was between 12.19 and 19.77 whereas for 1991-2001 it was between 0.18 and 4.55). However, when assessing the trend it is important to take into account the small values of the annual term frequency prior to 2001 (sometimes as low as 1).

The term frequencies and absolute term incidences of 'nanotechn*' in the title of articles on Academic Search Premier were calculated for three monthly intervals from 2001-03, and are presented

in Table 3. The reason for focusing on this period is that this is the period in which the absolute term incidences grew particularly rapidly.

Table 3: Term frequencies and absolute term incidences of 'nanotechn*' in the title (Academic Search Premier).

period	term frequency	absolute term incidence
Oct - Dec 2003	43	14.88
Jul – Sep 2003	42	14.99
Apr – Jun 2003	28	9.85
Jan – Mar 2003	34	13.15
Oct - Dec 2002	23	9.83
Jul – Sep 2002	16	6.95
Apr – Jun 2002	18	7.57
Jan – Mar 2002	13	5.72
Oct - Dec 2001	5	2.98
Jul – Sep 2001	5	3.05
Apr – Jun 2001	5	2.91
Jan – Mar 2001	7	3.94

The three-monthly absolute term incidences were substantially higher for the period Oct 2002-Dec 03 than for Jan-Sep 2002, which in turn was substantially higher than that for Jan-Dec 2001 (The three-monthly absolute term incidence for Oct 2002-Dec 03 was between 9.83 and 14.99, for Jan-Sep 2002 between 5.72 and 7.57, and for Jan-Dec 2001 between 2.91 and 3.94). However, when assessing the trend it is important to take into account the small values of the three-monthly term frequency (sometimes as low as 5).

1.3. Conclusion

In the Conclusion we briefly describe some the potential uses and limitations of absolute term incidences.

Absolute term incidences can be used to measure an enormous range of trends in articles held on Academic Search Premier and other searchable repositories of electronic text that provide data on the reported number of matches and on the number of items searched. However, in order for these trends to be useful it is important that they mirror trends of interest in the world outside the repository. In particular, for these trends in electronic text to be useful in research and science policy it is important that they mirror trends of interest in research and science policy.

We pose two questions: (1) To what extent do trends in electronic text mirror trends in the outside world or in research and science policy? and (2) What type of trends in the outside world and, in particular, in research and science policy are likely to be useful?

In order to answer the first question we find it useful to consider some of the limitations of using absolute term incidences to identify trends. These limitations include:

a) The growth of absolute term incidences could be a poor reflection of the growth of a concept. In the previous section we suggested the possibility that a high percentage of articles that are interdisciplinary or multidiscipliary in nature do not contain these words in the titles. If this is the case, then the growth of the absolute term incidences could be a poor reflection of the growth of the concept of disciplinarity.

b) Another limitation mentioned in the previous section is that sometimes the numbers on which the absolute term incidences are evaluated are so low that little significance can be attached to the trends. For instance, some of the absolute term incidences in Table 2 are calculated on quantities as small as 1. This limitation is likely to become less of a constraint, as the numbers on which the absolute term incidences are evaluated are generally higher for recent years that for a decade ago.

c) A third limitation is due to possible variations in the nature of the articles added to the repository. The reported number of articles on Academic Search Premier with date 2005 is more than twelve times that of 1988 (1,148,205 compared with 80,068). Although absolute term incidences adjust for changes in the number of items searched, they do not adjust for changes in the nature of the items.

d) The reported number of matches can be very inaccurate, in that it differs radically from the actual number of matches. We have applied some tests to identify inaccuracies in a diversity of repositories. Whilst we have not found any inaccuracies in the reported number of matches on Academic Search Premier, we have identified very high levels of inaccuracy in the reported number of matches on Google Scholar.

The second question is what types of trend are likely to be of use in the outside world and, in particular, in research and science policy. We suggest that monthly trends would seem more likely to be of use than yearly trends; however, repositories often are only searchable by year. As we are interested in focusing on trends that are likely to be of use, we would be very pleased to receive feedback on the trends that they consider to be in this category. Our email addresses are J.Levitt@ucl.ac.uk and JL794@tutor.open.ac.uk.

References

1. H. Small, Paradigms, citations, and maps of science: A personal history, Journal of the American Society for Information Science and Technology, 54 (5):394-399, American Society for Information Science and Technology, 2003.
2. T. Braun and A. Schubert, A quantitative view on the coming of age of interdisciplinarity in the sciences 1980-1999, Scientometrics, 58(1):183-189, Akadémiai Kiadó, co-published with Springer Science+Business Media B.V., 2003.
3. Y. N. Gu, Global knowledge management research: A bibliometric analysis, Scientometrics, 61(2):171-190, Akadémiai Kiadó, co-published with Springer Science+Business Media B.V., 2004.