on the internet – Do internet page counts provide latent indicators of scientific eminence?

Ludwig Maximilians Universität München

Andranik Tumasjan, Matthias Männich & Matthias Spörrle

ABSTRACT

Recently, Haggbloom et al. (2002) established a rank-ordered list of the 100 most eminent psychologists of the 20th century (though only the first 99 are actually reported by the authors) meticulously measured by several quantitative and qualitative indicators. We aimed at replicating this listing by simply using page counts obtained from three major internet search engines using different search queries with a five times repeated measurement. The resulting highly reliable indicators of internet frequency were consistently positively associated with the existing ranking and this correlation reached significance when the field of research was included in the query as an operator. We conclude that frequency data obtained by this method can be considered a simple and valid indicator of scientific impact and discuss additional applications of this method.

INTRODUCTION

Retrospectively appraising notable contributions to the field of psychology, Haggbloom et al. (2002) compiled an inclusive list of the 100 most eminent psychologists of the 20th century (though they report the first 99 only). Three qualitative and three qualitative indicators of eminence were used to establish the rank-ordered list. The quantitative indicators comprise the following measures: First, a journal citation frequency measure was used adding citation frequencies across four previously published citation lists. Second, an introductory textbook citation frequency list was constructed using three lists. Third, a survey was mailed to APA members asking to name the "greatest psychologists of the 20th century" in one's own opinion. The qualitative indicators used were the following dichotomous variables: (1) election to the NAS, (2) recipient of the APA Distinguished Scientific Contributions Award or elected APA president (as of 1999), or both, and (3) has his or her surname in use as an eponym. All six indicators were then used to calculate a composite score for each psychologist and resulted in the final rank-ordered list.

As the internet has become a major knowledge source where comprehensive information about the offline world is stored it seems reasonable to assume that popularity or influence of certain ideas (e.g., scientific impact) should also be reflected online. Therefore the aim of the present study was to replicate Haggbloom et al.'s (2002) rank-ordered list that was compiled on basis of offline resources by applying internet search engines' page counts as indicators of scientific eminence.

METHOD

Search engines:

We used three search engines to correct for possible biases and establish a more reliable indicator of internet frequency. The web services were chosen to meet two criteria: (1) Frequent usage within the Internet community and (2) high levels of independence from each other thus reducing the overlap of the search results. Three search engines met these criteria: www.google.com, www.yahoo.com and www.msn.com.

Queries:

Three types of query for the N = 99 psychologists' names were used: (1) "SN" {surname name} (e.g., Sigmund Freud), (2) "SN + psychology" {surname name psychology} (e.g., Sigmund Freud psychology), and (3) "SN + psychologist" {surname name psychologist} (e.g., Sigmund Freud psychologist). The queries were entered without further restrictions (e.g., no phrase search) to ensure that all websites where "surname", "name" and "psychology" or "psychologist" are mentioned together on one page (no matter in which conjunction) are taken into account. Otherwise results would have been limited to pages where the strings "surname name psychology/psychologist" are mentioned together in one phrase.

Procedure:

The query was done automatically by using a self-written internet spider. All queries were submitted at five times between April and May 2007.

RESULTS

Absolute frequencies (page counts) obtained were ranked in order to allow computation of Pearson correlations (for application of Pearson correlation coefficients on rank transformed data see Conover & Iman, 1981). Overall, the three internet search engines' rankings revealed high convergence across the three types of search query. All computed correlations between the search engines' rankings (pooled across the five points of measurement) for each type of query were between $r = .92$ and $r = .98$ (Table 1). Thus, all three search engines produced similar rankings for page count indicators. Moreover, the five times repeated measurements revealed a highly reliable rank order for all three search engines (Cronbach's alphas were between .99 and 1.00). Therefore all five measurements were pooled across the three search engines to form three overall internet rankings for the three types of search query. Subsequently, we computed Pearson correlation coefficients (cf. Conover & Iman, 1981) between Haggbloom et al.'s (2002) ranking and the rankings that resulted from the three types of query (Table 2). These results clearly show that our ranking based on internet page count indicators is significantly correlated with the offline ranking by Haggbloom et al. (2002).

Table 1: Pearson correlations between the three search engines for the three types of query

Type of query/ search engine	(1) "SN" MSN/Yahoo	(2) "SN + psychology" MSN/Yahoo	(3) "SN + psychologist" MSN/Yahoo
(1) "SN" Google	**.98**/.97****	.90**/.89**	.87**/.85**
(2) "SN + psychology" Google	.89**/.89**	**.94**/.93****	.92**/.93**
(3) "SN + psychologist" Google	.90**/.90**	.94**/.94**	**.92**/.93****

** $p < .01$

Table 2: Pearson correlations between Haggbloom et al.'s (2002) ranking and rankings for the three types of internet search query across all three search engines

Indicator	(1) "SN"	(2) "SN + psychology"	(3) "SN + psychologist"
Haggbloom et al.'s (2002) ranking	.12/.10/.10/**.10**	.26*/.22*/.18/**.20***	.29*/.26*/.18/**.21***

* $p < .05$; The first three correlation coefficients are for the rankings of (1) MSN, (2) Yahoo and (3) Google. The fourth coefficient represents the correlation with the mean ranking composed of all three search engines.

DISCUSSION

Results indicate that our ranking based on search engines' page counts and Haggbloom et al.'s (2002) offline ranking are positively and significantly correlated. Adding the field of research (i.e., psychology) or the profession (i.e., psychologist) to a scientist's name both result in a considerably higher correlation than querying surname and name only. All three search engines revealed high convergence in the rank order, thus a combined measure should provide a valid indicator. Furthermore, the present results suggest that internet page counts can be used as indicators for scientific impact. Further research should cross-validate this method replicating similar rankings and additionally using other search queries and/or operators for comparison.

References:

Conover, W. J. & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35,* 124-129.

Haggbloom et al. (2002). The 100 most eminent psychologists of the 20th century. *Review of General Psychology, 6*, 139-152.

Contact:

andranik.tumasjan@gmail.com, spoerrle@psy.uni-muenchen.de