THE UNIVERSITY of NORTH CAROLINA
GREENSBORO

Information about Information: Public Investments in Information Retrieval Research

Department of Economics Working Paper Series

Albert N. Link

University of North Carolina at Greensboro

Brent R. Rowe

RTI International

Dallas W. Wood

RTI International

April 2011
Working Paper 11-10

www.uncg.edu/bae/econ/

Information about Information:

Public Investments in Information Retrieval Research

Albert N. Link
Department of Economics
University of North Carolina at Greensboro
Greensboro, NC 27402
anlink@uncg.edu

Brent R. Rowe
RTI International
Research Triangle Park, NC 27709
browe@rti.org

Dallas W. Wood
RTI International
Research Triangle Park, NC 27709
dwood@rti.org

April 15, 2011

Abstract

Information retrieval (IR) is the science and practice of matching information seekers with the information being sought. Research on IR focuses on improving the effectiveness and efficiency of retrieval techniques and evaluating competing retrieval mechanisms. For example, Internet search engines utilize IR techniques to provide relevant information to users. In the United States, about $29 million of public support has been devoted to IR research over the past two decades. Through the activities of the Text Retrieval Conference (TREC) program with the U.S. National Institute of Standards and Technology (NIST). Here, we show empirically that research organizations worldwide that avail themselves of this information have relatively greater IR performance.

Keywords: Information retrieval, public goods, knowledge production function

Information about Information:
Public Investments in Information Retrieval Research

I. Introduction

Information retrieval (IR) is the science and practice of matching information seekers with the information being sought. IR techniques have been used to improve the process of finding information in other media ranging from desktop computers and to library databases to legal/medical records. For example, Internet users rely on IR-based tools when they rely on a Web search engine.

Research on IR focuses on improving the efficiency of retrieval techniques and evaluating competing retrieval mechanisms. This research takes place worldwide, in both government laboratories and in university laboratories. While the social impact of IR is pervasive, effecting individuals, firms, and public institutions, the paucity of previous research related to this information tool is surprising. The literature that exists focuses almost exclusively on describing technical improvements in performance of IR systems over time (Rowe et al. 2010).

This paper expands the scope of previous academic research by examining empirically factors associated with IR performance. Specifically, we investigate in a descriptive manner correlates with IR performance. Albeit and exploratory analysis, it is motivated with an eye toward public accountability. That is, the public sector in the United States has been involved in supporting infrastructures associated with IR research, but to date there has not been a systematic study of the effectiveness of those investments.

The paper is outlined as follows. In Section II, we set the stage for the remainder of the paper by briefly providing an overview of the evolution of post-world War II IR research in the United States. In Section III, we then describe U.S. public investments in IR research that support, in an infrastructural manner, such research. In Section IV, we posit a knowledge production function model suitable to investigate a relationship between the IR performance of IR research organizations and their level of use of public investments in IR. Finally, the paper concludes in Section V with summary remarks and a call for future research on this topic.

II. Origins of IR Research

In the aftermath of World War II there was a dramatic increase in the number of scientific articles being published, based in large part on the technological advances that the war engendered. As Vannevar Bush, Director of the U.S. Office of Scientific Research and Development, observed in response to this accumulating volume and the burgeoning need for computerized IR (1945, pp. 102, 107):

> There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends. … The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of the square-rigged ships. … Professionally our methods of transmitting and reviewing the results of research are generations old and by now are totally inadequate for their purpose. … Consider [as a solution] a future device for individual use, which is a sort of mechanized private file and library … in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility.

Shortly after the war, and partly in response to Bush's concerns, the Uniterm system was developed by Mortimer Taube of the U.S. Library of Congress. This system indexed documents based on key words rather than on subject headings as had been used over previous decades in, for example, card catalogues (Meadow et al. 2000).

Taube's efforts represented a major and unprecedented step forward. Prior to the 1950s, all of the tasks of an IR system were manual. Users seeking to identify documents related to particular topics relied on printed bibliographic indexes, generally available in the form of a library card catalogue. In the case of scientific literature, professional societies published indices from time to time to serve as a research guide (Meadow et al. 2000).

By the early 1950s, IR researchers seeking to improve or develop more efficient retrieval systems realized that matching documents using a key word index system, like Uniterm, was essentially a process that could be described by the algebra of sets and thus computers could understand the underlying algorithms. The first demonstration and implementation of such a computer-based search system was at the Naval Ordinance Test Station (NOTS) in California in 1954 using an IBM 701 computer (Bourne 1999).

As new computer-based indexing systems proliferated during the mid-1950s, the question of which index was most useful for IR was addressed by Cyril Cleverdon of Cranfield University in the United Kingdom (Robertson 2008). While his comparative study did not reveal significant differences among the existing computer-based systems in use at that time—the Uniterm system, the Universal Decimal Classification system, an alphabetical subject catalogue system, and a faceted classification system—it did establish a basic methodology for subsequent evaluation studies.[1]

New IR systems continued to be developed and they began to proliferate in public organizations. For example, in 1961 Gerald Salton wrote computer programs collectively known as the System for the Mechanical Analysis and Retrieval of Test (SMART). The programs used responses to sequential search queries to retrieve documents. By the mid-1960s, mainframe computers had sufficient speed and memory to perform IR searches in minutes rather than in hours. For example, in 1967, the National Library of Medicine contracted to have its Medical Literature Analysis and Retrieval System online (MEDLINE) accessed from terminals anywhere in the continental United States. Similar on-line systems proliferated throughout the rest of the 1960s and 1970s.

III. Public Investments in IR Research

During the late 1980s, the public sector undertook efforts to advance IR research because of the public good nature of knowledge *per se*. Market failure is the terse rationale for public investments in technology with a public or quasi-public good nature. More to the point, the public sector's role in innovation is based on an expected social return from its investments that is greater than society's hurdle rate, and on the private sector's inability to undertake the socially-desirable level of such investment.[2]

Donna Harman, an IR researcher at the National Institute of Standards and Technology (NIST), established using public-sector resources, the Citator System to illustrate the

[1] A detailed description of these systems is in Rowe et al. (2010).

[2] These arguments trace to the pioneering work of Arrow (1962) as articulated most recently by Link and Scott (2005, 2011).

effectiveness of statistical IR techniques on large test collections (Harman and Candela 1990). This initiative was followed by the creation of the Message Understanding Conferences (MUCs), a similar government effort initiated in the late 1980s by the Naval Command, Control and Ocean Surveillance Center (NOSC) and co-funded by the Defense Advanced Research Projects Agency (DARPA).

In 1992 the Text REtrieval Conference (TREC) program was established through collaboration between the National Institute of Standards and Technology (NIST), one of the national laboratories in the United States, and the Defense Applied Research Projects Agency (DARPA), the research office for the U.S. Department of Defense. The overall goal of TREC was to support and encourage research within the IR community by providing the infrastructure necessary for evaluation of IR methodologies using large data sets and to improve the transfer of IR technology from research laboratories to commercial products.[3]

TREC revolutionized IR system evaluation through five mechanisms: creation of new, larger test collections; development of standardized IR evaluation methods; establishment of an annual IR research workshop; a mechanism to distribute research results; and a model for other IR workshops.[4] By 1997, there was growing evidence that TREC had stimulated significant improvements in IR systems. Buckley et al. (1997) analyzed the performance of systems being evaluated and found that the majority of systems improvements between 1992 and 1997 could be attributed to knowledge and methods associated with TREC activities.

Such improvements have resulted from modest public investments. TREC investment costs, from 1991 through 2009 in real $2009, totaled $29 million (Rowe et al. 2010). Over time about 40 percent of total costs have come from NIST; about 40 percent from other government agencies including the National Security Agency and the Central Intelligence Agency; and about 20 percent from DARPA, mostly in the earlier years.

[3] As Tassey (2005), Link and Link (2009), and Link and Scott (2011) have argued, the provision of technology infrastructures like the Citator System and TREC are a responsibility of government. Although the perceived net social benefits are positive, research activities in which the private sector is highly uncertain of the appropriability of their investments (i.e., they may not capture all benefits) are unlikely to be funded with private resources.

[4] A detailed discussion of each of these TREC mechanisms is discussed in Rowe et al. (2010).

The remainder of this paper explores the relationship between improvements in IR system performance and infrastructural support to IR systems that are traceable to TREC's public investments in IR research.[5]

IV. Statistical Analysis

A. Description of the Data

In 2009, NIST commissioned a global survey of TREC researchers in an effort to understand the use of TREC resources by country and by research organizational type of participant/user (Rowe et al. 2010). As with many public-sector investments in technology, relevant supporting agencies examine their role retrospectively rather than prospectively (Tassey 2003).

The NIST database is used herein to examine correlates with improvements in IR performance. A total of 392 survey instruments were fully or partially completed; however, only 229 respondents reported both the country in which the surveyed organization resides and its research type. Table 1 shows the distribution of the 229 respondents by country and by organizational type. A total of 82 organizations are from the United States and 147 are from other countries. Over one-half (88) of the non-U.S. organizations are in Europe, 35 are in Asia and 24 are in non-European or Asian countries.[6]

Table 1 also shows the distribution of organizational type for U.S. versus non-U.S. organizations, as well as organizations from Europe and Asia. Over 75 percent of the respondents in the sample of 229 are from university or academic research laboratories. By country/continent, over 70 percent of U.S. organizations, nearly 74 percent of European organizations, and over 85 percent of Asian organizations were from university or academic research laboratories.

The focal variable in this analysis is IR performance, *PERFORM*. Respondents were asked two sequential questions:

[5] This research question is similar to that asked by Zucker et al. (2007) with respect to investments in nanotechnology, namely what are the factors that influence the rates at which new knowledge is produced in the nanotechnology field.

[6] Rowe et al. (2010) discuss data from 246 respondents, but that analysis includes organization types that were classified as Other. Respondents that did not cleanly fall in the categorical types in Table 1 are not considered in the analysis that follows. As well, Rowe et al. classified by continent some respondents outside of the United States based on their IP address. Those respondents were also not considered herein.

- *Did the quality of the IR system(s) developed by your research group improve during the time your organization used TREC projects and services?*
- *If yes, by how much (in percentage terms) did the quality of your IR system improve?*[7]

B. Empirical Model

We hypothesize that improvement in IR performance, *PERFORM*, is a function of the frequency that the unit attended TREC conferences and utilized TREC responses, measured in terms of the number of conferences attended and resources used, *UTILIZ.* Respondents were asked as part of the NIST survey to report the years in which they attended a TREC conference between 1992 and 2008, inclusive, and the years in which they used TREC resources. *UTILZ* is the sum of years attended and years of use.[8] In a very broad sense, our analysis represents a first-order effort to assess the performance consequences of public-sector infrastructural support through TREC.

We also hypothesize that the research size of the organizational unit, measured in terms of its IR R&D expenditures ($millions), *IRRD*, will have a positive impact on IR performance. The larger the organization, the greater the likelihood that the organization will realize economies of scale in research and thus greater IR performance. Respondents were asked to report their IR R&D expenditures within ranges (e.g., $100M to $249M). The mid-point of alternative ranges was used to quantify *IRRD.*

The framework for descriptive analysis is based on a knowledge production function, as popularized by Griliches (1979). A knowledge production function assumes that innovative output is a function of the stock of technical knowledge, among other resources, and has been posited to take a Cobb-Douglas function form. The stock of technical knowledge is then described as a function of investments in R&D. Within the context of our statistical analysis, the

[7] IR system improvement is generally based through an analysis of Mean Average Precision (MAP). A MAP score is calculated as follows. First, for a given query, the average of all the precision values is calculated at each recall point in a document. Then, the mean of all the query average precision scores is determined. The resulting MPA value provides a single measure of the quality of the IR program across recall levels (Manning et al. 2008). It is important to note that the percentage improvement in IR system performance may be related to the absolute maturity of the system. That is, if a researcher has just developed a new IR system, it may realize improvements of, say 50% to 75% over a period of time. However, a researcher who has been making gradual improvements to an IR system may only realize improvements that are one-tenth or less of this amount even though the latter system is of comparable quality, as measured by its MAP score. This vintage effect is controlled for, at least partially, through *IRRD* (see below).

[8] The theoretical range for *UTILZ* is 0 to 34.

variable *PERFORM* is a measure of innovative output, public investments into the stock of technical knowledge are represented by the variable *UTILIZ*,[9] and internal investments into the stock of technical knowledge are represented by the variable *IRRD*.

Thus, following Griliches (1979) and the academic research that followed:[10]

$$(1) \qquad PERFORM = \mathrm{F}(UTILIZ, IRRD) = UTILIZ^{\beta}\ IRRD^{\gamma}$$

For estimation purposes, equation (1) is written as:

$$(2) \qquad \ln PERFORM = \alpha + \beta\ UTILZ + \gamma\ IRRD + \varepsilon$$

where α is an intercept term and ε is a normal and randomly distributed error term. Based on the extant literature, one would expect β and γ to be positive; their level of significance is an empirical issue.

Also held constant in the estimable version of equation (2) are binary variables to control for whether the organization is U.S. or non-U.S.; or U.S., European, or Asian; and binary variables to control for whether the organization is an academic or government laboratory. We do not offer hypotheses for these control variables.

Descriptive statistics on all variables used in the estimation of equation (2) are in Table 2.

C. Regression Results

The least-squares regression results from equation (2) are reported in Table 3. The extent to which the research organization utilizes TREC resources, *UTLIZ*, is highly correlated with IR performance.[11] Broadly interpreted, this finding justifies government's involvement in IR research in the sense that it quantifies one dimension of the benefits attributable to such public investments.

Alternative versions of equation (2) were also estimated. In one set of equations *IRRD* is the research size regressor and in the other set *IRRD* and $IRRD^2$ are the regressors to account for

[9] Broadly speaking, the variable *UTILIZ* is a measure of each organizations investment in public education, and as Nelson and Phelps (1966) have shown such education is positively related to innovative output.

[10] The subsequent research has been summarized by Link and Siegel (2003).

[11] There is no statistical evidence that *UTLIZ* is non-linearly related to IR performance. These results are available on request from the authors.

the possibility of diminishing returns. Research size appears to also be linearly related to IR performance. The estimated coefficient on *IRRD* is marginally significant when the non-linear effect is controlled; see columns (2) and (4).

IR performance is not related to the geographical location of the research organization, where the organization is dichotomously measured in terms of U.S. versus non-U.S. location or in terms of U.S. versus Europe versus Asia versus the rest of the world.

Lastly, the IR performance of academic and government laboratories, holding laboratory size and utilization of TREC resources constant, is significantly greater than in IR service and software companies (subsumed in the intercept term).

V. Concluding Remarks

The findings presented in this paper, albeit exploratory in scope and structure, suggest that there are quantifiable social benefits associated with U.S. investments in IR research. Utilization of IR resources from TREC has a positive impact on the IR performance of the using research organization, other things held constant. As well, the finding that the knowledge associated with the use of TREC resources is not geographically bounded underscores not only the public good nature of knowledge *per se* but also it substantiates in part public investments in IR research.

Our findings should be interpreted with caution and no generalizations should be made about the impact of either public-sector investments in IR infrastructure or private R&D on IR knowledge.[12] More research is certainly needed not only from an evaluation perspective but also from an institutional perspective related to the breadth of importance of IR and it broad-based impacts on society.

[12] In addition, there was insufficient data—and theory about specific variables—to allow for a control for selection bias.

Table 1
Distribution of Respondents by Country and Organization Type

Country	Organization Type	Number of Respondents
US		
	IR Service and Software Company	18
	University or Academic Research Laboratory	58
	Government or Institutional Research Laboratory	6
		82
Non-U.S. (Europe, Asia, Other)		
	IR Service and Software Company	10
	University or Academic Research Laboratory	115
	Government or Institutional Research Laboratory	22
		147
Europe		
	IR Service and Software Company	8
	University or Academic Research Laboratory	65
	Government or Institutional Research Laboratory	15
		88
Asia		
	IR Service and Software Company	1
	University or Academic Research Laboratory	30
	Government or Institutional Research Laboratory	4
		35

Table 2
Descriptive Statistics for Variables Used to Estimate Equation (2)
n=122*

Variable	**Mean**	**Std Dev**	**Min**	**Max**
PERFORM	43.47	55.85	3.0	400.0
UTILIZ	9.95	7.85	0	32
IRRD	3.88	14.68	0.002	131.56
US	0.31	0.47	0	1
EUROPE	0.40	0.49	0	1
ASIA	0.18	0.39	0	1
ACADEMIC	0.76	0.43	0	1
GOVERNMENT	0.18	0.39	0	1

* Of the 229 survey respondents, only 122 reported all information used to estimate equation (2).

Table 3
Regression Results from Equation (2)
n=122
(standard errors)

Variable	(1)	(2)	(3)	(4)
UTILIZ	0.047	0.047	0.048	0.048
	(0.011)****	(0.011)****	(0.011)****	(0.012)****
IRRD	0.009	0.031	0.009	0.030
	(0.006)	(0.017)*	(0.006)	(0.018)*
$IRRD^2$	--	-0.0002	--	-0.0002
		(0.0002)		(0.0002)
US	-0.102	-0.106	0.119	0.099
	(0.196)	(0.195)	(0.307)	(0.306)
EUROPE	--	--	0.251	0.231
			(0.296)	(0.295)
ASIA	--	--	0.309	0.290
			(0.332)	(0.331)
ACADEMIC	0.634	0.710	0.622	0.696
	(0.369)*	(0.373)*	(0.371)	(0.375)*
GOVERNMENT	0.777	0.875	0.762	0.858
	(0.409)*	(0.415)**	(0.413)*	(0.419)**
constant	2.170	2.058	1.945	1.853
	(0.385)	(0.393)****	(0.454)****	(0.458)****
R^2	0.158	0.171	0.165	0.177
F-level	4.35***	3.95***	3.21***	3.03***

**** significant at .001 level, *** significant at .01 level, ** significant at .05 level, * significant at .10 level

References

Arrow, K.J. (1962). "Economic Welfare and the Allocation of Resources for Invention," in *The Rate and Direction of Inventive Activity*, Princeton: Princeton University Press.

Battelle, J. (2005). *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture,* New York: Penguin Group.

Bourne, C. (1999). "40 Years of Database Distribution and Use: An Overview and Observation," <http://www.nfais.org/publications/mc_lecture_1999.htm>.

Buckley, C., A. Singhal, and M. Mitra (1997). "Using Query Zoning and Correlation Within SMART: TREC 5," in *Proceedings of the Fifth Text Retrieval Conference*, NIST Special Publication, Gaithersburg, MD.

Bush, V. (1945). "As We May Think," *Atlantic Monthly* 176: 101-108.

Griliches, Z. (1979). "Issues in Assessing the Contribution of Research and Development to Productivity Growth," *The Bell Journal of Economics* 10: 92-116.

Harman, D. and G. Candela (1990). "Retrieving Records from a Gigabyte of Text on a Minicomputer Using Statistical Ranking," *Journal of the American Society for Information Science* 41: 581-589.

Link, A.N. and J.R. Link (2009). *Government as Entrepreneur*, New York: Oxford University Press.

Link, A.N. and J.T. Scott (2005). *Evaluating Public Research Institutions*, London: Routledge.

Link, A.N and J.T. Scott (2011). *Public Goods, Public Gains: Calculating the Social Benefits of Public R&D,* New York: Oxford University Press.

Link, A.N. and D.S. Siegel (2003). *Technological Change and Economic Performance*, London: Routledge.

Manning, C.D., P. Raghavan, and H. Schütze (2008). *Introduction to Information Retrieval,* New York: Cambridge University Press.

Meadow, C., B. Boyce, and D. Kraft (2000). *Text Information Retrieval Systems,* San Diego: Academic Press.

Nelson, R.R. and E.S. Phelps (1966). "Investment in Humans, Technological Diffusion, and Economic Growth," *American Economic Review* 56: 69-75

Robertson, S. (2008). “On the History of Evaluation in IR,” *Journal of Information Science* 34: 439-456.

Rowe, B., D. Wood, A.N. Link, and D. Simoni (2010). *Economic Impact Assessment of NIST’s Text REtrieval Conference (TREC) Program*, final report submitted to the Program Office, National Institute of Standards and Technology, Gaithersburg, MD.

Tassey, G. (2003). “Methods for Assessing the Economic Impacts of Government R&D,” NIST Planning Report 03-1, Gaithersburg, MD.

Tassey, G. (2005). “Underinvestment in Public Good Technologies,” *Journal of Technology Transfer*, 30: 89-113.

Zucker, L.G., M.R. Darby, J. Furner, R.C. Liu, and H. Ma (2007). “Knowledge Stocks, Knowledge Flows and New Knowledge Production,” *Research Policy* 36: 850–863.