A Retrospective and Prospective View of Information Retrieval and Artificial Intelligence in the 21st Century

Eugene Garfield
Chairman Emeritus, ISI®, Publisher, The Scientist®, 3501 Market Street, Philadelphia, PA 19104. E-mail: garfield@CODEX.cis.upenn.edu; Home Page: http://garfield.library.upenn.edu

Introduction

In the 1950s, scientists typically scanned a dozen or so journals they personally received. They visited libraries to cover the rest of the literature. In those days, journals were much more affordable. A survey of the American Chemical Society chapter in Philadelphia found that most members subscribed to more journals than they read. It was the time of the reprint culture (Garfield, 1999). Authors exchanged reprints generously. It was not unusual to mail reprints regularly to members of one's invisible college. The now ubiquitous Xerox machine was not available, and photostats were expensive and cumbersome.

Correspondence by snail mail was the norm, as was the use of printed indexes and abstracting services. The pace of research and the publication process was significantly slower. Once published, however, the time to deliver journals and reprints, including transatlantic steamship delivery was remarkable. Rudi Schmid described the speed of intercontinental and transcontinental mail transit times from Europe to California from 1852 to 1941 (Schmid, 1984). The transit and mail postal system is now considered archaic. Nevertheless, most print journals still use snail mail for domestic distribution augmented by air-cargo services for international distribution. After World War II, the introduction of telephone, fax, and then e-mail created a completely new situation.

On-line access to indexing and abstracting services was introduced in the 1970s. Two decades later, full-text journal articles on-line began to appear and now are routinely provided. The integration of the journal literature with A&I services through linking services presents a completely transformed situation. Readers are now instantly accessing journal contents pages, abstracts, cited references with abstracts, and full text. It is possible to browse the current literature on-line and in real time go backward and then forward again into related documents. As full-text archives increase their chronological scope, it will be possible to search and peruse the literature without ever entering the library.

Within 5 years, scientists will be able to access most of the last 10 years of the literature electronically. In a decade, this will extend to much of the journal literature of the 20th century, especially for the 1000 most-consulted journals. Full conversion will depend upon the cost of scanning back runs of journals, following the JSTOR model. However, Dana Roth (Roth, 1999), rejecting the JSTOR model, has suggested we create files of the most-cited papers. Although the highest impact journals of science are currently available, electronically complete archives are still a rarity. The 500 most-cited journals identified by ISI's *Journal Citation Reports*® are listed on Highwire's Web site. But there are still formidable barriers unless your library has an electronic site license for all these journals.

An alternative interim step is to use e-mail to contact authors rapidly for access to articles not yet directly available on the Web. Some articles may even be found on the author's personal home page. It would greatly improve the situation if each institution assumed responsibility for creating digital libraries of the articles produced by their faculty, especially those who are retired or deceased. It would be equally helpful if university Web sites provided a standardized means of access to faculty email addresses.

The creation of large digital libraries seems inevitable, especially if technology continues to reduce the cost of conversion from paper. Large-scale conversions to PDF files are possible at a cost of about 50 cents to $1.00 per page. Projects like JSTOR are intended to take care of the archiving gap even as individual authors self-archive. Clearly, there is a tacit desire to archive everything that has been published. However, a situation that is half-electronic and half-paper will inevitably lead to equally half-baked retrospective coverage of the literature. Authors take the path of least resistance. Obtaining anything not archived on the Web is increasingly costly.

Searching Full Text

While alerting and SDI services were available 40 years ago, it is now rather routine for societies and publishers to announce forthcoming articles electronically. The time lag between submission and publication of articles is rapidly diminishing, as is the work of preparing and editing manuscripts. The need to standardize formats for electronic documents is evident, as is the desire to standardize electronic manuscripts per se. One can rely on the services of Pro-Cite, End-Note, Reference Manager, or other database management systems to produce articles in any journal style required without having to completely retype the manuscript. Overall, these systems have increased the efficiency of producing or reformatting original manuscripts. Furthermore, the increased use of personal Web pages displaces the need to go directly to the library for a lot of archival material. Like other authors, I have "self-archived" most of what I have published in my career. Lenhoff has recently suggested that retired scholars do this systematically, especially the work they have not published (Lenhoff, 2000).

Even the time-consuming process of peer review seems to have been accelerated, because electronic access facilitates the paperwork involved. E-mail receipt of manuscripts provides a stimulus to potential referees to act promptly and make it very inexpensive to increase the pool of referees.

Sending Full Texts

Searching full texts of documents presents new and interesting problems.

Information scientists have been studying full-text searching for 50 years. John O'Connor was one of the pioneers. Early on he recognized the need to create artificially intelligent searching systems (O'Connor, 1965). Personal experience with large-scale files, including even my own, demonstrates the blessing and the dilemmas of full-text searching. For the rare word or phrase, it is extremely efficient. For the frequently occurring term, it can be highly frustrating. Twentieth-first-century users will demand more sophisticated methods for refining such searches.

The speed of access to electronic files is an important factor in our ability to take advantage of full-text scanning. The ability to display groups of documents rapidly for scanning and weeding is essential to the process of information recovery. I experience the elation and frustration of full-text when I use the Verity system to search my own publications. The full-text is available on-line. However, to take full advantage of its word-for-word indexing, I need to be able to instantly pop up the context in which the term occurs, not just the title of the article. Such systems need to display the context, as is demonstrated in the autonomous citation index developed by Lawrence et al. (1999).

SDI Profiling and Clipping Services

A few years ago, I wrote the following letter to the *New York Times* about push–pull technology and its predecessors. The letter was not published (Garfield, 1997).

> In her article on ". . .how I came to hate push technology," (*The New York Times*, p. C5, March 24, 1997) Denise Caruso speculates whether "Push Technology" signals the doom of the Web browser (Caruso, 1997). However, on March 23, 1997, in "Pushy, Pushy," *New York Times Magazine* (p. 32), James Gleick provides a cogent response (Gleick, 1997). My experience with "Push Pull" technology may be of interest.
>
> In 1965 Irving H. Sher and I created *Research Alert* (Garfield & Sher, 1967), the first commercially available computer-based system for selective dissemination of information (SDI). Since then the service has been operated continuously on a weekly basis by the *Institute for Scientific Information* (*ISI*). The key to its success is timing, comprehensiveness, and high degree of specificity. Since the early seventies *DIALOG*, *Lexis-Nexis*, and other on-line systems have also provided "Push Pull" technology. The success of SDI services is based on their highly selective profiling systems. Unless "Push Technology" or current Web "crawlers" do the same, they will frustrate most users. Significantly improved search engines will make Web browsers increasingly valuable, even while equally improved SDI (Push) systems gain popularity.
>
> *Pointcast* and other broadly based systems are relatively useless to most users but they can become highly specific, as they are with individual stocks.
>
> The needs of scientists, medical researchers, and scholars are quite varied and only systems that can provide the ability to customize literature searching will be used repeatedly. Broader dissemination is provided by such tools as *Current Contents* and *Medline*, and hundreds of leading specialty journals.
>
> Profiling systems are widely used in the information industry to follow patent, journal, and other literature. The level of specificity needed often involves complex combinations of descriptors, but also the ability to identify current publications that quote specific papers and people.
>
> Existing Web "crawlers" do not provide an acceptable level of precision and convenience, but competition will force them to rediscover what the library and information community has known for over three decades.

The *ASCA* system developed by Irv Sher, myself, and others at ISI is often described as SDI—selective dissemination of information—a special kind of current awareness. Clipping services have existed since the beginning of the last century, but the ISI personal alert system (*ASCA*) for the first time dealt with the huge body of scientific and scholarly literature (Garfield & Sher, 1967).

Thirty-five years after launching the *Automatic Subject Citation Alert*, it is difficult to estimate the extent to which SDI is used. I see minimal evidence of this in academia. Certain institutions like Stanford have made it popular by using the ISI database in combination with SDI software developed by Los Alamos National Laboratory. Information professionals have an important educational task to make users "profile" conscious so that they will embrace these "push–pull" systems. In particular, they must learn to take full advantage of keyword and citation profiling. While not called citation profiling, this capability has been incorpo-

rated in the Highwire Press system (online at http://highwire.stanford.edu). For each new article one encounters, the user can automatically include its citation as part of an alerting profile.

Foreign Language Translation

A significant amount of literature is still published in foreign languages. The ability to use translation dictionaries facilitates the ability to read foreign language material. Using pop-up windows to translate individual words or phrases, much as one uses a spell checker, can be extremely time saving. Given a real-time word-for-word look-up system, I can read most papers in German, Spanish, or French with minimum difficulty. As I have pointed out recently (Garfield, in press), a great deal of editorial comment is still expressed in vernacular languages, so this translation capability is important to those who wish to take into account the opinions of foreign authors. Foreign editors should take advantage of these translation facilities to produce multilingual versions of their editorials and articles. Because the translations can appear on journal Web sites, the cost of publishing multilingual versions can be significantly reduced (Watters & Patel, 1999). Systran (www.systran.com) is one such system that often does a remarkable job of "quick and dirty" translation but does not yet provide the convenience of quick pop-up word-for-word translation as is done with RichLink Technology at www.babylon.com or www.sentius.com.

Information Nirvana

In the early days of my career, I referred to an information nirvana (Garfield, 1962). This is yet another metaphor for the World Brain of H.G. Wells and the dreams of the early encyclopedists. Each new generation of information technology advancements brings with it a need for new refinements. The notion of the automatic review of the literature has been in the minds of information scientists for a long time. Whether we can ever obtain artificially intelligent machines for creating reviews, remains to be seen. Displaying lists of citations surrounded by contextual text is just one obvious step (Small, 1978).

Research scientists, especially in the life sciences, need to parse scientific documents so that key phrases used in various combinations can lead to interesting correlations. Sher used phrase analysis to create Keywords Plus (Garfield, 1990; 1990; Garfield & Sher, 1993). This sort of parsing is common to computational linguistic programs. It is unlikely that automation can replace the human intelligence necessary to make these correlations. It is possible to imagine that these new systems of artificial intelligence will facilitate the indexing needed in fields like evaluative medicine or bioinformatics. The pharmaceutical and biotechnology industries are now dependent upon a whole new subindustry involving structure–function determination and correlation.

The prototype for this type of aposteriori intelligence is John O'Connor's brilliant attempt to develop systems for scanning the full text of a document, which never mentions the word toxic or toxicity and yet an intelligent automation could conclude that it contains an indication of toxicity (O'Connor, 1965).

Another expression of the AI challenge is implicit in the distinction I made in 1965 (Garfield, 1965) between an automated system of citation indexing such as autonomous citation indexing (Lawrence, 1999), and a system that is able to read a text and supply the missing references (Watters & Patel, 1999). The experiment that I personally conducted with a group of graduate students, demonstrated that the need for a cited reference in a text is perceived quite differently depending upon the reader's sophistication. Given an article I had published in the *Journal of Chemical Documentation* (Garfield, 1961), students were asked to insert a mark wherever they thought a reference was needed. The number of references varied from 15 to 75, but averaged about 35, which, in fact, was close to what I had used (Garfield, 1977).

From the preceding remarks, it will not be surprising that I hold in high esteem the work of Don Swanson in attempting to create an artificially intelligent agent for generating correlations between disease elements and potential therapies (Swanson & Smalheiser, 1997; 1999).

All such experiments emphasize the unique role played by the critical review in the progress of science. This role is needed increasingly even as we gain easier access to the primary literature. It is the *a posteriori* use of the literature that paves the way to discovery. That is what the IR game is all about. Information systems should facilitate the process of making new connections. In the meantime, human, mainly laboratory-based researchers, continue this creative process of reviewing. Organizations like *Annual Reviews*, *Current Science*, and others already provide a rich supply of such reviews. The huge output of review articles and their high impact demonstrates, I believe, their value to the scientific community. Twenty years ago, ISI and *Annual Reviews* established the National Academy of Sciences Award in Recognition of this role (Garfield, 1979).

Perhaps the most significant advance in reviewing has been made by the Cochrane Collaboration Centers (http://www.cochrane.de/), which form the basis for modern evidence-based medicine. The success of that enterprise may now be applied to other problems based on the formation of the new Campbell Collaboration (http://campbell.gse.upenn.edu/). Electronic journals and databases will aid these systems of synthesis, but should significantly reduce publication bias because space in printed journals will not be a limiting factor (Song et al., 1999).

Information Discovery and Recovery

This leads to a concluding observation. Information retrieval concerns both information discovery and information recovery (Garfield, 1969; 1966). While closely related, the

process of information recovery should approach perfection in the years to come.

We should rarely have difficulty in recovering papers we have encountered in the past. Information discovery systems; however, will remain a daunting challenge for decades to come because they involve the injection of human intelligence difficult to match in AI systems. Recognizing how long it has taken to reach the present state of the art, I doubt that many of us will still be here when these breakthroughs occur.

References

Babylon.com, http://www.babylon.com

Campbell Collaboration, http://campbell.gse.upenn.edu/.

Caruso, D. (1997). Push technology. New York Times, p. C5, March 24, 1997.

Cochrane Collaboration Centers, http://www.cochrane.de/.

Garfield, E. (1961). Information theory and other quantitative factors in code design for document card systems. Journal of Chemical Documentation, 1(1), 70–75. Reprinted 1977 in Current Contents, 44, 8–19, and in 1980 in Essays of an information scientist, vol. 3 (pp. 274–285), Philadelphia: ISI Press. http://www.garfield.library.upenn.edu/essays/v3p274y1977-78.pdf.

Garfield, E. (1962). The ideal library—The informatorium. Current Contents, 1. Reprinted 1977 in Essays of an Information Scientist, vol. 1. Philadelphia: ISI Press. http://www.garfield.library.upenn.edu/essays/V1p001y1962-73.pdf.

Garfield, E. (1963). Research nirvana—Total dissemination and retrieval of bio-medical information? Paper presented at Sixth Annual Session, Medical Writers' Institute, New York City, October 5.

Garfield, E. (1965). Can citation indexing be automated? In M.E. Stevens, V.E. Giuliano, & L.B. Helprin (Eds.), Statistical association methods for mechanized documentation, symposium proceedings, 1964 (pp. 189–192). National Bureau of Standards Miscellaneous Publication 269. http://www.garfield.library.upenn.edu/essays/V1p084y1962-73.pdf.

Garfield, E. (1966). ISI eases scientists' information problems; Provides convenient orderly access to literature. Karger Gazette, 13, 2. Reprinted in 1969 as The who and why of ISI. Current Contents, 13, 5–6, which was reprinted in 1977 in Essays of an information scientist, vol. 1 (pp. 33–37), Philadelphia: ISI Press. http://garfield.library.upenn.edu/essays/V1p033y1962

Garfield, E. (1969). ISI's comprehensive system of information services. Current Contents, 3. Reprinted in 1977 in Essays of an information scientist, vol. 1 (p. 31), Philadelphia: ISI Press. http://garfield.library.upenn.edu/essays/V1p031y1962-73.pdf.

Garfield, E. (1977). Information theory and all that jazz: A lost reference list leads to a pragmatic assignment for students. Current Contents, 44, 5–19. Reprinted in 1980 in Essays of an information scientist, vol. 3 (pp. 271–273), Philadelphia: ISI Press. http://www.garfield.library.upenn.edu/essays/v3p272y1977-78.pdf.

Garfield, E. (1979). The NAS James Murray Luck award for excellence in scientific reviewing: G. Alan Robison receives the first award for his work on cyclic AMP. Current Contents, 18, 5–9. Reprinted in 1981 in Essays of an information scientist, vol. 4 (pp.127–131), Philadelphia: ISI Press. http://www.garfield.library.upenn.edu/essays/v4p127y1979-80.pdf.

Garfield, E. (1990). KeyWords Plus: ISI's break-through retrieval method. Part 1. Expanding your searching power on current contents on diskette. Current Contents, 32, 3–7 Reprinted 1991 in Essays of an Information Scientist, vol. 13 (pp. 295–299), Philadelphia: ISI Press. http://www.garfield.library.upenn.edu/essays/v13p295y1990.pdf.

Garfield, E. (1990). KeyWords Plus takes you beyond title words. Part 2. Expanded journal coverage for current contents on diskette includes social and behavioral sciences. Current Contents, 33, 5–9. Reprinted 1991 in Essays of an Information Scientist, vol. 13. Philadelphia: ISI Press. http://www.garfield.library.upenn.edu/essays/v13p300y1990.pdf.

Garfield, E. (1997). SDI and "push pull" technology. Letter to the Editor, New York Times, April 14, 1997 (unpublished). http://www.garfield.library.upenn.edu/papers/pushpull.html.

Garfield, E. (1999). Evolution of the reprint culture: From photostats to home pages on the World Wide Web: A tutorial on how to create your electronic archive. The Scientist, 13(4), 14. http://www.the-scientist.library.upenn.edu/yr1999/feb/comm_990215.html.

Garfield, E. (in press). Foreign journal editors should translate their editorials on the Web. The Scientist. ** See url below

Garfield, E., & Sher, I.H. (1967). ASCA (automatic subject citation alert)–A new personalized current awareness service for scientists. American Behavioral Scientist, 10(5), 29–32. http://www.garfield.library.upenn.edu/essays/v6p514y1983.pdf. Reprinted 1984 in Essays of an Information Scientist, vol. 6 (pp. 514–517), Philadelphia, PA: ISI Press.

Garfield, E., & Sher, I.H. (1993). KeyWords Plus—Algorithmic derivative indexing. Journal of the American Society for Information Science, 44(5), 298–299. http://www.garfield.library.upenn.edu/papers/jasis44(5)p298y1993.html.

Gleick, J. (1997). Pushy, pushy. New York Times Magazine, p. 32, March 23, 1997.

Highwire Press, http://highwire.stanford.edu/.

Lawrence, S. (1999). Digital libraries and autonomous citation indexing. Computer, 32(6), 67. http://www.neci.nec.com/-lawrence/aci.html.

Lenhoff, H.M. (2000). An e-journal for a vanishing resource. The Scientist, 14(2), 35. http://www.the-scientist.com/yr2000/jan/opin_000124.html.

O'Connor, J. (1965). Automatic subject recognition in scientific papers—An empirical study. Journal of the ACM, 12(4), 490.

Roth, D. (1999). Private communication.

Schmid, R. (1984). Naturae novitates, 1879–1944—Its publication and intercontinental transit times mirror European history. Taxon, 33(4), 636–654.

Sentius Corporation, http://www.sentius.com

Small, H. (1978). Cited documents as concept symbols. Social Studies of Science, 8, 317–340.

Song, F., Eastwood, A., Gilbody, S., & Duley, L. (1999). The role of electronic journals in reducing publication bias. Medical Informatics and the Internet, 24(3), 223–229.

Swanson, D.R., & Smalheiser, N.R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. Artificial Intelligence, 91(2), 183–203.

Swanson, D.R., & Smalheiser, N.R. (1999). Implicit text linkages between medline records: Using Arrowsmith as an aid to scientific discovery. Library Trends, 48(1), 48–59.

Systran Corporation, http://www.systran.com

Watters, P.A., & Patel, M. (1999). Semantic processing performance of internet machine translation systems. Internet Research–Electronic Networking Applications and Policy, 9(2), 153–160.

** http://www.garfield.library.upenn.edu/commentaries/tsv14(09)p06y20000501.pdf - Published in The Scientst 14[9]:6, May 1, 2000