Undiscovered Public Knowledge: a Ten-Year Update

Don R. Swanson* and Neil R. Smalheiser†

*Division of Humanities, University of Chicago,
1010 E. 59th St.,Chicago, IL 60637; swanson@kiwi.uchicago.edu
†Department of Pediatrics, University of Chicago,
5841 S. Maryland Ave., Chicago, IL 60637; sma2@midway.uchicago.edu

Abstract

Two literatures or sets of articles are complementary if, considered together, they can reveal useful information of scientific interest not apparent in either of the two sets alone. Of particular interest are complementary literatures that are also mutually isolated and noninteractive (they do not cite each other and are not co-cited). In that case, the intriguing possibility arises that the information gained by combining them is novel. During the past decade, we have identified seven examples of complementary noninteractive structures in the biomedical literature. Each structure led to a novel, plausible, and testable hypothesis that, in several cases, was subsequently corroborated by medical researchers through clinical or laboratory investigation. We have also developed, tested, and described a systematic, computer-aided approach to finding and identifying complementary noninteractive literatures.

Specialization, Fragmentation, and a Connection Explosion

By some obscure spontaneous process scientists have responded to the growth of science by organizing their work into specialties, thus permitting each individual to focus on a small part of the total literature. Specialties that grow too large tend to divide into subspecialties that have their own literatures which, by a process of repeated splitting, maintain more or less fixed and manageable size. As the total literature grows, the number of specialties, but not in general the size of each, increases (Kochen, 1963; Swanson, 1990c).

But the unintended consequence of specialization is fragmentation. By dividing up the pie, the potential relationships among its pieces tend to be neglected. Although scientific literature cannot, in the long run, grow disproportionately to the growth of the communities and resources that produce it, combinations of implicitly-related segments of literature can grow much faster than the literature itself and can readily exceed the capacity of the community to identify and assimilate such relatedness (Swanson, 1993). The significance of the "information explosion" thus may lie not in an explosion of quantity per se, but in an incalculably greater combinatorial explosion of unnoticed and unintended logical connections.

The Significance of Complementary Noninteractive Literatures

If two literatures each of substantial size are linked by arguments that they respectively put forward -- that is, are "logically" related, or complementary -- one would expect to gain useful information by combining them. For example, suppose that one (biomedical) literature establishes that some environmental factor A influences certain internal physiological conditions and a second literature establishes that these same physiological changes influence the course of disease C. Presumably, then, anyone who reads both literatures could conclude that factor A might influence disease C. Under such conditions of complementarity one would also expect the two literatures to refer to each other. If, however, the two literatures were developed independently of one another, the logical linkage illustrated may be both unintended and unnoticed. To detect such mutual isolation, we examine the citation pattern. If two literatures are "noninteractive" — that is, if they have never (or seldom) been cited together, and if neither cites the other, — then it is possible that scientists have not previously considered both literatures together, and so it is possible that no one is aware of the implicit A-C connection. The two conditions, complementarity and noninteraction, describe a model structure that shows how useful information can remain undiscovered even though its components consist of public knowledge (Swanson, 1987, 1991).

Public Knowledge / Private Knowledge

There is, of course, no way to know in any particular case whether the possibility of an AC relationship in the above model has or has not occurred to someone, or whether or not anyone has actually considered the two literatures on A and C together, a private matter that necessarily remains conjectural. However, our argument is based only on determining whether there is any printed evidence to the contrary. We are concerned with public rather than

private knowledge -- with the state of the record produced rather than the state of mind of the producers (Swanson, 1990d). The point of bringing together the AB and BC literatures, in any event, is not to "prove" an AC linkage, (by considering only transitive relationships) but rather to call attention to an apparently unnoticed association that may be worth investigating. In principle any chain of scientific, including analogic, reasoning in which different links appear in noninteractive literatures may lead to the discovery of new interesting connections.

"What people know" is a common understanding of what is meant by "knowledge". If taken in this subjective sense, the idea of "knowledge discovery" could mean merely that someone discovered something they hadn't known before. Our focus in the present paper is on a second sense of the word "knowledge", a meaning associated with the *products* of human intellectual activity, as encoded in the public record, rather than with the contents of the human mind. This abstract world of human-created "objective" knowledge is open to exploration and discovery, for it can contain territory that is subjectively unknown to anyone (Popper, 1972). Our work is directed toward the discovery of scientifically-useful information implicit in the public record, but not previously made explicit. The problem we address concerns structures within the scientific literature, not within the mind.

The Process of Finding Complementary Noninteractive Literatures

During the past ten years, we have pursued three goals: i) to show in principle how new knowledge might be gained by synthesizing logically- related noninteractive literatures; ii) to demonstrate that such structures do exist, at least within the biomedical literature; and iii) to develop a systematic process for finding them.

In pursuit of goal iii, we have created interactive software and database search strategies that can facilitate the discovery of complementary structures in the published literature of science. The universe or search-space under consideration is limited only by the coverage of the major scientific databases, though we have focused primarily on the biomedical field and the MEDLINE database (8 million records). In 1991, a systematic approach to finding complementary structures was outlined and became a point of departure for software development (Swanson, 1991). The system that has now taken shape is based on a 3-way interaction between computer software, bibliographic databases, and a human operator. The interaction generates information structures that are used heuristically to guide the search for promising complementary literatures.

The user of the system begins by choosing a question or problem area of scientific interest that can be associated with a literature, C. Elsewhere we describe and evaluate experimental computer software, which we call ARROWSMITH (Swanson & Smalheiser, 1997), that performs two separate functions that can be used independently. The first function produces a list of candidates for a second literature, A, complementary to C, from which the user can select one candidate (at a time) as an input, along with C, to the second function. This first function can be considered as a computer-assisted process of problem-discovery, an issue identified in the AI literature (Langley, et al., 1987; p304-307). Alternatively, the user may wish to identify a second literature, A, as a conjecture or hypothesis generated independently of the computer-produced list of candidates.

Our approach has been based on the use of article titles as a guide to identifying complementary literatures. As indicated above, our point of departure for the second function is a tentative scientific hypothesis associated with two literatures, A and C. A title-word search of MEDLINE is used to create two local computer title-files associated with A and C, respectively. These files are used as input to the ARROWSMITH software, which then produces a list of all words common to the two sets of titles, except for words excluded by an extensive stoplist (presently about 5000 words). The resulting list of words provides the basis for identifying title-word pathways that might provide clues to the presence of complementary arguments within the literatures corresponding to A and C. The output of this procedure is a structured title-display (plus journal citation), that serves as a heuristic aid to identifying word-linked titles and serves also as an organized guide to the literature.

Seven Examples of Literature-Based Knowledge Synthesis

The concept of "undiscovered public knowledge" based on complementary noninteractive literatures was introduced, developed, and exemplified in (Swanson, 1986a, 1986b). Since 1986, we have described six more examples, each representing a synthesis of two complementary literatures in which biomedical relationships not previously noted in print were brought to light. We describe also the hypotheses to which they have led, and the strategies we have followed in finding and identifying these structures (Swanson 1988, 1989a, 1989b, 1990a; Smalheiser & Swanson 1994, 1996). We identify these examples here in terms of A, B, and C, wherein associations between A and B are found in one literature and associations between B and C in another literature, leading us to draw certain inferences about a previously unreported association between A and C. In most cases we analyzed multiple B-terms for a given A and C. In the

following description we identify only A and C, and a few of the more important B-connections, along with the main conclusion or hypothesis to which we were led. Other authors have reviewed, extended, or assessed this work (Chen, 1993; Davies, 1989; Garfield, 1994; Gordon & Lindsay, 1996; Lesk, 1991; Spasser, in press).

Example 1, 1986: Dietary Fish Oils (A) and Raynaud's Disease (C)

Dietary fish oils (esp. eicosapentaenoic acid) lead to certain blood and vascular changes (B) that are separately known to be beneficial to patients with Raynaud's disease. One B-linkage, for example, was: dietary eicosapentaenoic acid can decrease *blood viscosity* (B); abnormally high *blood viscosity* has been reported in patients with Raynaud's disease. (Swanson, 1986a, 1986b, 1987). The inference that fish oils may benefit Raynaud patients may be regarded as a successful prediction; two years after publication of the above analysis, the first clinical trial demonstrating such a beneficial effect of fish oil was reported by medical researchers (cited and discussed in Swanson, 1993).

Example 2, 1988: Magnesium Deficiency (A) and Migraine (C).

B consists of eleven indirect connections, which led to a prediction that magnesium deficiency might be implicated in migraine headache. (Two such linkages are, for example: magnesium can inhibit *spreading depression* in the cortex, and *spreading depression* may be implicated in migraine attacks; magnesium-deficient rats have been used as a model of *epilepsy*, and *epilepsy* has been associated with migraine) (Swanson, 1988, 1989b). Since publication of that analysis, more than 12 different groups of medical researchers have reported a systemic or local magnesium deficiency in migraine or a favorable response of migraine patients to dietary supplementation with magnesium (cited and discussed in Swanson, 1993).

Example 3, 1990: Arginine (A) and Somatomedin C (C).

B consists of five physiologic associations leading to the inference that orally administered arginine may increase blood levels of somatomedin C (the latter being known to have a number of beneficial effects). For example, infused arginine stimulates the release of growth hormone, and the latter in turn is known to increase blood levels of somatomedin C (Swanson, 1990a). Our inferences led to a proposal by medical researchers to conduct a clinical trial (discussed further in (Swanson, 1993)).

Example 4, 1994: Dietary Magnesium (A) and Neurologic Disease (C).

Endogenous magnesium ions play a key role in regulating excitotoxicity mediated by the NMDA receptor (B). Excitotoxicity in turn is thought to have an important role in various neurologic diseases. We suggested that the possible effect of manipulating exogenous (e.g. dietary) magnesium on brain function or neurologic disease merits investigation (Smalheiser & Swanson, 1994).

Example 5, 1995: Indomethacin (A) and Alzheimer's Disease (C)

In this example, the A and C literatures are neither disjoint nor noninteractive. Indeed, there is clinical and epidemiologic evidence that indomethacin may have a protective effect against Alzheimer's disease. However, we found certain indirect associations (B) between the two literatures that were not mentioned in the direct (A-C) literature, or elsewhere. One of these B-relationships in particular indicated that indomethacin, because of its anti-cholinergic activity, could adversely affect Alzheimer patients by exacerbating cognitive dysfunction. Because this possibility apparently had not been previously reported, we brought it to the attention of neuroscientists (Smalheiser & Swanson, 1996).

Example 6, 1995: Estrogen (A) and Alzheimer's Disease (C)

As in example 5, the A and C literatures are interactive; estrogen replacement therapy is reported to be associated with a lower incidence of Alzheimer's disease, but the mechanism of such an effect is unknown. We reported several previously-uninvestigated B-relationships, particularly one involving antioxidant activity, that appeared to merit investigation as possible explanations of this intriguing relationship (Smalheiser & Swanson, in press).

Example 7, 1996: Phospholipases (A) and Sleep (C)

The A and C literatures are disjoint and noninteractive, but implicitly related through a set of substances (notably interleukin 1=DF, tumor necrosis factor and endotoxin/lipopolysaccharide) which are well known both to promote sleep and to stimulate one or more phospholipases. This study identified a list of agents whose effects on sleep are especially likely to involve phospholipases; and suggested several straightforward experimental tests of our hypothesis that phospholipases may be involved in endogenous pathways that regulate sleep (Smalheiser & Swanson, in preparation).

Comment

The objects of study in the work summarized here are complementary structures within the scientific literature. The recognition of meaningful associations and ultimately that of complementarity require a high level of subject expertise. The unruly problems of meaning within the natural language of titles and abstracts present serious obstacles to more fully automating this process of knowledge discovery. Our computer aids are therefore designed to enhance and stimulate human ability to see connections and relationships. These aids necessarily derive from the immense databases that provide the routes of intellectual access to the literature. Our goal thus far has been to produce a working practical system that yields immediate results in furthering the aims of biomedical research, and which at the same time generates data and problems that contribute to understanding literature-based scientific discovery.

References

Chen Z. 1993. Let documents talk to each other: A computer model for connection of short documents, *The Journal of Documentation* 49(1):44-54.

Davies, R. 1989. The Creation of New Knowledge by Information Retrieval and Classification. *The Journal of Documentation* 45(4):273-301.

Garfield, E. 1994. Linking literatures: An intriguing use of the Citation Index, *Current Contents* #21, 3-5.

Gordon, M.D. & Lindsay, R. K. 1996. Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature based discovery of a connection between Raynaud's and fish oil. *Journal of the American Society for Information Science* 47:116-128.

Kochen, M. 1963. On natural information systems: pragmatic aspects of information retrieval. *Methods of Information in Medicine* 2(4):143-147.

Langley, P., Simon, H. A., Bradshaw, G. L., and Zytkow, J. M. 1987. *Scientific Discovery. Computational Explorations of the Creative Process.* Cambridge, Mass.: MIT Press

Lesk, M. 1991. SIGIR '91: The More Things Change, the More They Stay the Same. In: *SIGIR FORUM.* 25(2):4-7, ACM Press.

Popper, K. R. 1972. *Objective Knowledge* Oxford.

Smalheiser, N.R. and Swanson, D. R. 1994. Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. *Neurosci Res Commun* 15(1):1-9.

Smalheiser, N.R. and Swanson, D. R. 1996. Indomethacin and Alzheimer's Disease. *Neurology* 46:583.

Smalheiser, N.R. and Swanson, D. R. (in press) Linking Estrogen to Alzheimer's Disease: An informatics approach. *Neurology*

Spasser, M. (in press) The enacted fate of undiscovered public knowledge. *Journal of the American Society for Information Science.*

Swanson, D. R. 1986a Undiscovered public knowledge. *Library Quarterly* 56(2):103-118.

Swanson, D. R. 1986b. Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge. *Perspectives in Biology and Medicine* 30(1):7-18.

Swanson, D. R. 1987. Two Medical Literatures that are Logically but not Bibliographically Connected. *Journal of the American Society for Information Science* 38(4):228-233.

Swanson, D. R. 1988. Migraine and Magnesium: Eleven Neglected Connections. *Perspectives in Biology and Medicine* 31(4):526-557.

Swanson, D. R. 1989a. Online Search for Logically-Related Noninteractive Medical Literatures: A Systematic Trial-and-Error Strategy. *Journal of the American Society for Information Science*40:356-358.

Swanson, D. R. 1989b. A Second Example of Mutually-Isolated Medical Literatures Related by Implicit, Unnoticed Connections. *Journal of the American Society for Information Science* 40:432-435.

Swanson, D. R. 1990a. Somatomedin C and Arginine; Implicit Connections Between Mutually-Isolated Literatures. *Perspectives in Biology and Medicine* 33(2):157-186.

Swanson, D. R. 1990b. Medical Literature as a Potential Source of New Knowledge. *Bulletin of the Medical Library Association* 78(1):29-37.

Swanson, D. R. 1990c. Integrative Mechanisms in the Growth of Knowledge: A legacy of Manfred Kochen. *Information Processing & Management* 26(1):9-16.

Swanson, D. R. 1990d. The absence of co-citation as a clue to undiscovered causal connections, in Borgman, C. L., ed. *Scholarly Communication and Bibliometrics.* 129-137. Newbury Park, CA: Sage Publ.

Swanson, D. R. 1991. Complementary Structures in Disjoint Science Literatures. In *SIGIR91 Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* Chicago, Oct 13-16, 1991 ed. A. Bookstein, et. al. New York: ACM; p. 280-9.

Swanson, D. R. 1993. Intervening in the Life Cycles of Scientific Knowledge, *Library Trends* 41(4):606-631.

Swanson, D. R. and Smalheiser, N. R. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Forthcoming.