Don R. Swanson
Impact on Information Science

Summary

Don R. Swanson (1924- 2012) pioneered the field of literature-based discovery, which uses existing research to create new knowledge. With a background as a computer systems analyst working on the relationship between natural and computer languages, he believed electronic databases to be essential to helping information science keep up with the increasing demands for more effective and efficient information services. His key works in demonstrating the effectiveness of literature-based discovery to find previously unknown connections among existing scientific literature and developing the ARROWSMITH system (with Neil Smalheiser) launched an entire field of study within information science.

For his contributions, he was awarded the ASIS&T’s highest honor, the Award of Merit, in 2000.

About

Swanson studied physics, receiving his M.A. (1947) and Ph.D. (1952) from Rice and the University of California, Berkeley, respectively. He then worked as a computer systems analyst and later a research scientist before becoming dean of the University of Chicago Graduate Library School in 1963. He served three terms as dean before retiring as professor emeritus in
the Humanities Division. His work continued up until his death, with an editor publishing his incomplete paper in 2011 for its historical and noteworthy value. (Swanson, 2011.)

**Computer-Aided Information Retrieval**

Believing that computers systems could automate and aid information retrieval, he brought a systematic and logical approach to the challenges he saw facing information science. Swanson began studying the problem of scholarly communication and retrieval and took both a philosophical and systematic approach to his analysis and proposed solutions. His early papers discussed the problem of the volume of information and ability to search for it (Swanson, 1960, 1966, 1977) which can be summarized as:

1. The volume of documents was rapidly increasing.
2. The audience consuming these documents was growing and was increasingly from outside the traditional target audience for a particular discipline
3. The increasing specialization in science was increasing the isolation of groups of researchers and their documents reducing the effectiveness of the formal scholarly communication system to support cross-discipline communication.
4. The then current information services’ and systems’ ability to consume, organize, retrieve and disseminate the volume of documentation was inadequate to meet this new state of information needs. He points to the rise of informal communications to work around this inability to keep up with this change.
He also questioned the means of measuring relevance. That is, how does one know that all relevant documents have been retrieved without having read every document in existence? In fact, he joins others in questioning “relevance” as a measurement for information retrieval systems as assessing “relevance” is dependent on the requestor. In many of his earlier papers, he takes a philosophical approach in his discussion and may not even touch on computer systems. However, he take a very systematic and logical approach to his analysis of information retrieval, such that when the technology became available, the information services professions would be organized and prepared to take advantage of it. (Swanson, 1977).

His more technical papers proposing methods for automatic indexing and retrieval still did not mention specific computer systems; but the suggested methodology that those systems should use (Bookstein, Swanson, 1974). Perhaps this was a strategic tactic for introducing the use of computer-aided information retrieval, which was still a novel concept at the time.

“Undiscovered Public Knowledge”

Swanson was concerned that the increased specialization and isolation of scientific disciplines and research and the lack of formal communication support for “cross-specialty” communications would lead to more missed opportunities to make connections between literatures.

Swanson believed that unearthing unseen links between two distinct areas of study could yield new discoveries—what he called “undiscovered public knowledge.” He wanted to demonstrate the presence of these “undiscovered connections” between two different sets of
disjoint literatures; that is, the two sets do not cite each other and show little evidence that the researchers in one group were aware of the other and develop a computer-assisted, literature-based approach to scientific discovery.

He approached this problem by asking if two separately published articles one stating "A may cause or influence B" and the other "B may cause or influence C" aren't both known to the same person or group of people; how can the conclusion be reached that A may cause or influence C? How can the literature be retrieved and presented in a way to make the connection between A and C apparent to the different people interested in A and C?

“Fish Oil and Reynaud’s Disease”

Using the existing medical literature, Swanson tested his theory of undiscovered public knowledge by hypothesizing a connection between dietary fish oil and Reynaud’s Syndrome. Raynaud’s Syndrome is a circulatory disorder commonly noted by platelet aggregability, high blood viscosity, and vasoconstriction. Concurrently, dietary fish oil has been shown to lead to reductions in platelet aggregability, blood viscosity, and vascular reactivity. These reductions have been shown to improve blood circulation. Is it possible to show through the literature that fish oil could be used in treating Reynaud’s Syndrome sufficiently that a researcher could justify a lab experiment?

First, Swanson collected an estimated 2000 papers on Reynaud’s Syndrome and 1000 on dietary fish oil over the previous decade. Then he demonstrated that both set of documents were disjoint with no citations, co-authorships, or connections between them. Then, by examining the titles and abstracts, he extracted main attributes (keywords, terms, phrases) in both sets and
found they had many in common (blood viscosity, platelet aggregability and vascular reactivity) that were related to the hypothesis he was testing.

Swanson published two papers as a result of this study (Swanson 1986a, 1986b). In *Library Quarterly* (Swanson, 1986b), he discussed the problem of undiscovered knowledge in disjoint literature from a philosophical perspective and challenged the library science field to take on the work. He described the challenge in information retrieval of uncovering “all relevant” documents would in fact necessitate directly examining every document that has ever been published. He suggested that computers systems and databases could assist humans in this information retrieval action. The second paper (Swanson, 1986a) was published in *Perspectives in Biology and Medicine*, where he digs into the literatures and demonstrates with an example how the connection could not be easily discovered because the literatures were so disjoint. He explained the logical inference of the hypothesis and ends with the proposed hypothesis.

Swanson continued applying his literature-based discovery process (Swanson, 1987, 1988, 1990) demonstrating that complementary but disjoint, non-interactive structures in the literature of science do exist and can lead to novel scientific hypotheses that are well worth testing. Three years after publishing his groundbreaking work (Swanson, 1986a), a clinical trial validated the use of fish oil for patients with Raynaud’s disease (DiGiacomo, 1989). He later hypothesized a connection between migraine headaches and magnesium deficiency (Swanson, 1989) that was also subsequently supported by clinical research. Based on these successes, he started to devise search strategies to uncover previously unknown connections in medical literature which he called “partially systematic”. The computer would assist with the tedious
tasks of indexing and retrieving the relevant literature; but a human was needed to guide the
discovery by determining the queries and assessing the relevancy of retrieved connections.

**Migraine – Magnesium Methodology**

In (Swanson, 1989) he differed his strategy from the beginning in that he started with the
topic of migraine; but did not have a secondary topic to connect it to. Migraine had no known
cause or cure and the connection to magnesium was unknown. Swanson, himself, did not start
with that hypothesized connection whereas with Fish-Oil and Reynaud’s Syndrome, he did. In
this case, he wanted to let the unknown connections appear. *Figure 1*, below, illustrates his
search process for discovery.

---

*Figure 1. Image reproduced from (Swanson, 1989)*

---
He started by retrieving the literature on migraine by searching online for titles that contain the term, approximately 4600 articles. Then he manually examined the titles looking for terms that might suggest a possible mechanism for migraine. Next he used these terms for further searching (not necessarily including the term migraine) and collected an intermediary literature that could contain links to possible cures and treatments. For example (which can also be seen in Figure 1) the first search returned terms including “spreading depression.” After retrieving literature on “spreading depression”, Swanson learned that “magnesium in the extracellular cerebral fluid can prevent or terminate spreading depression” (Swanson, 1989). He hypothesized that magnesium might alleviate migraine. Again, this hypothesis was later validated in clinical lab studies.

The significance of this second study demonstrated, as did the Fish-Oil and Reynaud’s study, that a partially systematic method for detecting undiscovered relationships within the science literature was viable and a worthwhile endeavor. Of course, whether or not the particular hypotheses proposed were correct was beside the point. Later validation of proposed hypotheses, however, was noted and prompted continued development of his methodology. Swanson’s next move was to determine if this search strategy could be more automated, widely applied, and reproduced successfully.

ARROWSMITH

As previously mentioned, the two previous studies were partially systematic in that Swanson performed an online database search to retrieve the literatures; yet still needed to manually scan the titles and select important terms. Swanson joined with Neil Smalheiser of the
University of Illinois at Chicago to build a “systematic, computational” method to find possible links among articles. “The computer was not supposed to generate discoveries, but it was supposed to identify and put together these potential assertions.” (Smalheiser, Swanson, 1998). They designed a set of interactive software and database search strategies, collectively called ARROWSMITH. ARROWSMITH is modeled little differently from the partially systematic method where a researcher performed one search, then mined keywords from the results to perform a new search. In this model, the searcher retrieves two unconnected literature sets, uploads them into ARROWSMITH and the system locates the connections and inferences. Besides supplying the two literature sets, a human still reviews and evaluates the intermediary steps to refine ARROWSMITH’s findings.

*Figure 2*, illustrates a basic ARROWSMITH application where the researcher starts with two topics that are unconnected; but suspects that connections can be revealed. Two literature sets are retrieved from MEDLINE, “A” and “C”, and both are uploaded into ARROWSMITH. ARROWSMITH locates frequent and unique terms of interest from both sets and the researcher evaluates each list, “AB” and “BC,” and manually edits each and removes non-relevant terms. Then ARROWSMITH presents the researcher with two lists of titles who compares the titles to evaluate a possible relation.
Figure 2 Image reproduced from (Smalheiser, Swanson, 1988)
Swanson’s work continues to influence Literature-based Discovery research to this day. Google Scholar report that his paper on Fish Oil and Reynaud’s Syndrome has been cited 665 times and Migraine and Magnesium 359 times. Both papers continue to be cited up though 2014, with the Migraine and Magnesium paper cited once so far in 2015. Robert Kostoff used the ARROWSMITH system to develop his methodology for what he now terms Literature Related Discovery and Innovation (Kostoff, 2012). More recently, a team attempted to apply text-mining techniques to further automate the manual evaluation performed by humans (Spangler & Wilkins, 2014) citing Swanson’s early work.

Note that Swanson’s literature-based discovery methodology ends after proposing possible hypotheses for further examination by the appropriate domain researchers. Swanson believed that finding these connections was well within the realm of information science; however he cautioned that information scientists’ goal should be to produce new hypotheses or suggestions – not discoveries. (Even though the sub-field is called “Literature-based Discovery”). Lab scientists can then test the hypotheses as their view is that the real discoveries happen in the lab, not the literature. This may have been politically motivated not wanting to step on other domains’ turf and lay claim to any “discoveries” eventually proven in the lab; but with Swanson’s background as a physical scientist, he was well aware that the connections still had to be verified in the lab to be considered. In his 2000 ASIS&T award acceptance speech, he advised the information science community, “Our job is to assemble other people's ideas” (Swanson, 2001).
References


Swanson, D. R. (1986a). Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, 30(1), 7-18.


