

Literature-Based Discovery? The Very Idea

D.R. Swanson

Abstract How is it possible to extract new knowledge from something already published? The possibility arises, for example, when two articles considered together suggest information of scientific interest not apparent from either article alone. In that sense, the two articles are complementary, a relationship based on the scientific problems, findings, and arguments presented. Whether the information found is also new and can lead to a plausible, testable hypothesis requires further searching and analysis of the literature from which it emerged.

The purpose of this introduction is to outline goals, concepts, problems, and literature structures that offer one approach to understanding the potential and limitations of literature-based discovery (LBD) independently of specific computer techniques that may be used to assist or implement it. The seeds of most of the basic concepts of LBD can be seen within the following classic exemplar of complementarity from a century ago that was of extraordinary importance to science.

1 Complementarity of Hybridity and Cytology: The Birth of Cytogenetics

A classic work by Walter Sutton in 1903 represents a landmark in genetics known as the Boveri–Sutton hypothesis [5]. The famous 1866 paper by Mendel on pea hybridization, that resurfaced in 1900, was interpreted by Sutton in the light of chromosome behavior observed in cell division and fertilization. An introduction to Sutton’s article, written by Peters in 1959, bears a remarkable resemblance to literature-based discovery:

“... When an author takes a series of apparently unrelated facts and ideas from two areas of investigation, combines them so that they make new sense, and develops a new hypothesis

D.R. Swanson

University of Chicago, 1010 E. 59th Street, Chicago, IL 60637, USA
swanson@uchicago.edu

P. Bruza and M. Weeber (eds.), *Literature-based Discovery*,
Springer Series in Information Science and Knowledge Management 15.
© Springer-Verlag Berlin Heidelberg 2008

from the combination, he not only aids in the advance of both fields but also is quite likely to open up a new one... In Sutton's paper you will see this development of relationships between the fields of cytology and heredity, which, at the time Sutton wrote, were considered to be fairly divergent from one another, in that no research techniques were shared... Sutton's paper can be considered the beginning of cytogenetics... This paper is a good model to follow in the preparation of a study involving synthesis and correlation." [5, p. 27]

To get some idea of the nature of the complementarity in Sutton's synthesis, a few salient features of Mendel's paper on hybridity and of the separate work in cytology may be helpful [8].

Mendel experimentally bred strains of peas with distinctive visible traits and found that hybrids from parents that breed true and differ consistently in one trait all look like the parent with the "dominant" trait. If those hybrids are then inbred, their first generation descendants show a 3:1 ratio of dominant to "recessive", where the latter is a reversion to the non-dominant grandparent. Continuing through one more generation, Mendel found that the recessives do not further vary, but only 2/3 of those bred from the hybrids that possess the dominant character show again a 3:1 ratio, thus indicating that the original 3:1 ratio could be reinterpreted as 1:2:1. Mendel continued his experiments through about six generations, and concluded that his results could be explained by assuming that the dominant and recessive traits split up within the new seeds and pollen, and then recombine at random during fertilization. He went on to show that a second pair of traits behaved in exactly the same way and as though entirely independent of the first pair.

Mendel's experiments entailed more than 10,000 plantings of peas. A cartoon shows a group of monks sitting at a dining table, and one monk walking in with a huge serving bowl. The caption reads: "Brother Mendel, we are getting a little tired of peas".

Turning now to the separate field of cytology (as of 1900) its basic data were derived not from plant breeding experiments but rather from examining the cell nucleus using a microscope. In the process of germ-cell division, paired chromosomes, one from each parent, separate to form gametes. Observations of the detailed behavior and orientation of chromosomes prior to dividing led Sutton to suggest that the gametes formed are just as likely to receive any given chromosome from one parent as the other. He saw that if he associated a pair of parental traits with a pair of parental chromosomes, then he could account for Mendel's observation that traits appeared to split up within the germ cells and recombine at random during subsequent fertilization. It also appeared that a second pair of traits and chromosomes behaved independently of the first pair, and so could account for Mendel's laws of segregation and distribution. Thus the problem posed by Mendel, of how a pair of traits can behave as though they were randomly distributed to progeny, is solved by Sutton's interpretation of chromosome behavior during meiosis and fertilization, which provides a causal mechanism sufficient to explain Mendel's results.

Sutton's paper holds at least two important lessons for literature-based discovery.

First, the two fields of experimental hybridity and cytology of the cell nucleus were good prospects for the analysis of complementarity even prior to 1900 because they were addressed to a common problem, in this case the transmission of hereditary traits.

Second, a more detailed study of cytology, focused on cell division and fertilization during the two decades before 1900, suggests that Sutton's synthesis was far more than a mechanistic process of putting two things together. It involved both inventiveness and substantial knowledge, both implicit and explicit. Once Sutton, who was a cytologist, had become aware of the Mendel paper, even a supercomputer of today would have been of little use in helping him create or interpret connections between cytology and hybridity. Recognizing complementarity is quintessentially a human function that requires creativity, inventiveness, scientific knowledge, and background knowledge – the latter including commonplace knowledge such as is needed for, among other things, understanding natural language in scientific text, or any other text, or to understand the point of a metaphor, a joke, or a cartoon, all of which depend on usage, context, and situation [2].

I know of no reason to believe that Sutton's achievement, notwithstanding its extraordinary importance, is unique in its dependence on human mental abilities in order to recognize complementary relationships. The achievement presents a challenge to people who understandably want the computer to do most of the work. Indeed, it seems likely that the creativity required is not unrelated to the importance of Sutton's work. If we design LBD procedures to find important connections by stimulating human creativity, the less important will follow by default, but not vice versa.

The goal of LBD in my opinion should be to support and enhance human ability by focusing on the key problems of finding promising pairs of scientific articles that can serve as a stimulus, and on identifying associated literature structures (see below). It is, in any event, plausible to assume that two articles randomly selected from a vast literature would have almost no chance of being complementary, so we need a search process that combines human knowledge and judgment with computer speed and data capacities. One key problem here is to determine what kind of clues are helpful in pointing to or defining "promising" pairs of articles.

2 Suggestive Complementarity and the ABC Model

"Complementarity", as defined above, is only suggestive, rather than compelling because scientific arguments expressed in natural language seldom lend themselves to logical description, largely because the background knowledge necessary for transforming the text of an article into a logical statement is almost always missing and typically taken for granted.

However, many scientific arguments are expressed as an association between two or more entities – such as substances and diseases in the biomedical literature. The idea of combining two entities is useful in providing a structured example of complementarity. One article might argue that term A is associated with B, and a second article that B is associated with C, in the absence of any explicit published claim that A may be related to C. This structure resembles a syllogism, but "association" and "relatedness" are not transitive so one should not be misled by the resemblance. I shall try to show that it is nonetheless useful for explanatory purposes. An AC

relationship under the circumstance given would be implicitly suggested and so worth thinking about to any reader who understands both articles (A, C), a key point being the word “suggestive”. Assuming that the two articles have no authors in common, it is also of interest to notice that the suggestion of an A–C relationship is unintended by the authors of either A or C who may not even have been aware of each other’s work.

The ABC model, even though overly simple, is sufficiently rich to serve as a useful point of departure and as a vehicle for an organized approach to defining literature structures that have a good chance of being relevant to LBD. Moreover, the A to B to C structure can be described also by the algebra of sets, wherein we consider the set of all articles containing term A and similarly for B and for C. AB and BC are then defined in terms of set intersections. To form and combine sets of articles is the function of the core search commands for the major bibliographic databases that provide routes of intellectual access to the literature of science.

Gardner-Medwin, an eminent biomedical researcher, presciently observed in 1981: “In these days of library computers it is possible to search the literature for papers linking two or more keywords. If one were to pick out the following associations neuroglia–potassium; potassium–spreading depression; spreading depression–migraine, one would make quite an impressive collection. Try to link neuroglia with migraine however, and there would be little to show. The aim of this paper is to explore the three associations set out above” [3, 7].

Gardner-Medwin proposed and executed a core idea of LBD in a single paper with the help of a computer database search, but otherwise without benefit of computer assistance. This approach can be seen as the ABC model extended to ABCD, and was published 4 years before the work on LBD was initiated in information science, where it was called “undiscovered public knowledge” [6]. Even though the Gardner-Medwin article received about 100 citations (up to April 2007), mostly on spreading cortical depression (a neurological brain phenomenon), only six articles turned up in a Medline search on neuroglia ‘AND’ migraine (in the title or abstract or as medical subject headings), none earlier than 1981. Three of these cited Gardner-Medwin. Unlike the spectacular impact of the 1903 Sutton synthesis, the Gardner-Medwin hypothesis does not appear to have stimulated much further research on neuroglia and migraine, even though the intermediate steps of the connection were argued in depth, frequently cited, and well-researched. I could find no further published work by Gardner-Medwin along the line of ABC-type connections.

3 What People “Know” versus Recorded Knowledge

It is important, in the context of what is meant by “novelty” to distinguish between what people “know” (or think they know), and what is published. Literature-based discovery is concerned not with state of mind but rather with the state of the public record. It follows that the novelty of any implicit discovery hinges not on whether

one or more scientists previously knew about it, but only on whether it had been previously made explicit in published form.

The journal article is one of the most important inventions supporting the infrastructure of modern science, dating from the mid-eighteenth century [10, 11]. Its function is to represent a small fragment of science, relatively short and to the point, that can then serve as a “building block” available for public use in a communal effort to construct the edifice of scientific knowledge, a process in which the blocks themselves may evolve into more mature forms and interact with their neighbors to form literatures addressed to common problems.

The size of the recorded knowledge base is far beyond human capacity to assimilate, even with the division of labor that specialization makes possible. And to include implicit knowledge based on connections increases the disparity enormously. Concerning human capacity, there is perhaps one exceptional case that has been reported:

He is the master of Balliol College What he doesn't know just isn't knowledge [1, p. 190].

4 Fragmentation of Science

The concept of LBD arises from and depends on three essential and interlocking aspects of recorded scientific knowledge – its immense size, an attempt to cope with size by specialization, and the resulting inevitable fragmentation of science into insular communities.

Specialization in science began along with the scientific journal. The patterns of communication, particularly in citation practices, are difficult to analyze prior to the era of Eugene Garfield and the citation indexes, but manual techniques with limited objectives are not infeasible. Hybridity and cytology between 1866 and 1900, the period during which Mendel's paper was reputedly neglected, is worth a closer look for our purposes.

Whether in fact the Mendel paper was neglected and, if so, why, has occasioned much published debate, but, more to the point for LBD is the paucity of published citations by cytologists to any of the hybridity literature (including Mendel) and vice versa, even though the two fields did share a common interest in the problem of hereditary transmission, at least after 1881. I was able to find, after substantial manual searching, only a few isolated examples of cross citations between the two fields, but these did not lead to significant or ongoing interaction. Judging from the citation pattern, both cytologists and hybridists seemed to be fully occupied within their own specialties, no doubt because that is why specialties developed in the first place. Sutton's breakthrough of the cytology–hybridity boundary in 1903 seemed to be virtually unprecedented.

Most of the LBD work to date has been based on the literature of biology and medicine, perhaps because the biological world is so richly interconnected. There are many scientific bibliographic databases, but the largest two in biology

and medicine illustrate the immense size of the literature today, with about 16 million articles covered by Medline and 18 million by BIOSIS (Biological Abstracts) (with substantial overlap of the two databases). Both of these databases are well-organized, indexed in depth, and associated with powerful, flexible search languages. They are the preeminent routes of access to the recorded knowledge of biology and medicine. The size of this vast literature necessarily shapes the nature of problems that LBD addresses.

Fragmentation is also manifest in the growth of the published record. Specialties do not tend to grow so large as to be unmanageable; prior to that point, subspecialties are formed spontaneously. Subspecialties therefore proliferate while maintaining a more or less limited maximum size of each. The literature of science cannot grow faster than the communities that produce it, but not so with connections. Implicit connections between subspecialties grow combinatorially. LBD is challenged more by a connection explosion than by an information explosion.

5 A Problem-Oriented Approach

The various approaches to research on LBD involve in one way or another some combination of human and machine procedures. Here, in order to bring into focus underlying principles, I envision LBD primarily as a human function, but in need of computer assistance for individual biomedical researchers.

A reasonable start for individual users of an LBD system is to define a problem in their own field of research and on that basis design a customized approach appropriate to solving the problem. The creation of relevant sub-literatures – principally by conducting searches using bibliographic databases, as did Gardner-Medwin, – is of great, perhaps overriding, importance to defining problems of manageable size.

A distinction between closed vs open ended searching is relevant in this context. Any LBD search that does not begin by clearly specifying a problem can be doubly open-ended, having, like the universe, neither a beginning nor an end [1, pp. 169–171]. Wishing to avoid questions of either cosmology or theology, I prefer to assume that one always starts with a user-defined problem that anchors the beginning of a search. The terminus is then open or closed. The open terminus often may be decomposed into multiple termini defined by a list of candidate terms suggested by either a human or a computer procedure. Any single choice from the list then characterizes a closed-end search, which is necessarily a hypothesis, not an established or confirmed finding. It remains to be tested in the laboratory, clinic, or other contexts in the real world, in the usual manner of scientific investigation.

The approach described above is individualized in that it envisions that LBD serves, and is used by, a subject specialist (e.g., a biomedical scientist) engaged in research. This approach encourages a focus on what can be done now to produce scientifically acceptable and useful results. Individualized small scale trial-and-error procedures are characterized by many dead ends and a few promising paths. We can learn from both failures and successes to develop requirements and techniques for future systems. Such an approach based on dispersed knowledge and exploratory searching is conducive to evolutionary improvement.

6 Complementary but Disjoint Structures in the Literature of Science

To determine whether supposedly new information seen in a pair of complementary articles has been published explicitly elsewhere – i.e., is not really new in terms of the state of the published record, requires a thorough literature search that may be far from a straightforward exercise.

The concept of novelty is domain dependent [8]. If we were to choose the world-wide domain of all recorded knowledge, it is impossible to prove that something is novel – i.e. does not exist elsewhere. Information retrieval is, in essence, an incomplete and uncertain process [2, 6, p. 113]. Yet, for all practical purposes a limitation to the major bibliographic and citation databases, and a high-recall search, would seem to be a reasonable basis for determining whether a connection is new to the published record, at least until proved otherwise.

The definition of complementarity can in an obvious way be extended from a pair of articles to a pair of sets of articles with each set characterized by substantially the same scientific argument. The question of whether the two sets intersect is then crucial. The new information that one might hope to gain from bringing together complementary individual articles may well already be contained in any overlapping set. In short, two complementary sets that have any substantial number of articles in common are probably of little interest for LBD.

Moreover it would be unusual for two sets of articles that cite each other extensively to be disjoint – i.e. have no articles in common, so for practical purposes it is reasonable and easier to determine the intersection of the two sets, and then only in the case of small or null intersections, check also for any citations from one literature to the other. In this context, normally one would expect two disjoint clusters to be unrelated and not complementary, and two complementary clusters to overlap extensively.

The foregoing argument suggests that two sets of articles that are complementary but disjoint (CBD) would represent an unusual structure – but it is just such a structure that commands the highest interest for LBD and is or should be the prime focus of LBD research, because the implicit results of complementary relationships that can be seen or deduced are probably undocumented and hence novel. They are likely also to be unknown and unintended [8].

The concept of disjoint, as used in CBD, is an idealization not to be taken rigidly. If relatively few articles are within the intersection of two much larger sets (say A, C), few enough so that it is not too difficult to directly examine each one to assess whether it represents a biologically meaningful connection between A and C, then for practical purposes A and C are disjoint. For any intersection paper that does represent a valid connection, the citation pattern can reveal whether or not it has been neglected, as may have been the case for Mendel's paper. LBD then might play a key role in strengthening and updating the literature-based connections (by analyzing more B-terms), and so calling attention to any neglected discovery that it might represent. I have given an example of such literature-based resurrection in a recent publication [9, pp. 1088, 1091].

7 Summing Up

I have suggested one way of thinking about literature-based discovery, stressing the point that understanding goals, problems, and concepts should precede consideration of how computers can be used to best advantage. LBD originates with the scientist as user defining a problem of interest and then examining combinations of articles that together suggest new hypotheses not apparent in the separate articles. These combinations are to be found in complementary but disjoint (CBD) literatures, the process of recognizing complementarity depending on human ingenuity. CBD literatures are formed by searching the major bibliographic databases, beginning with a user-defined problem and appropriate search strategies. The goal of an LBD system should be to stimulate human creativity in order to produce a plausible and testable hypothesis stated in a form suitable for publication in the subject field studied, where it is then open to testing, criticism, review, and stimulation of further research.

7.1 Postscript: A Warning About Consequences

In connection with a procedure very like LBD, a serious adverse effect has been predicted:

“...some day the piecing together of dissociated knowledge will open up such terrifying vistas of reality,...that we shall either go mad from the revelation or flee from the deadly light into the peace and safety of a new dark age.” – Lovecraft [1, 4].

Acknowledgements I thank Neil Smalheiser for valuable suggestions.

References

1. H.P. Barrow. *Impossibilities, The Limits of Science and the Science of Limits*. Oxford University Press, Oxford, 1998
2. D. Blair. *Wittgenstein, Language and Information: Back to the Rough Ground*. Springer, Berlin Heidelberg New York, 2006. [Part III]
3. A.R. Gardner-Medwin. Possible roles of vertebrate neuroglia in potassium dynamics, spreading depression, and migraine. *Journal of Experimental Biology*, 95:111–127, 1981
4. H.P. Lovecraft. The Call of Cthulhu. In S.T. Joshi, editor. *The Call of Cthulhu and Other Weird Stories*. Penguin Books Ltd., London, 1999
5. W.S. Sutton. The chromosomes in heredity. In J.A. Peters, editor, *Classic Papers in Genetics*, pp. 27–41. Prentice Hall, Englewood Cliffs, NJ, 1959
6. D.R. Swanson. Undiscovered public knowledge. *Library Quarterly*, 56:103–118, 1986
7. D.R. Swanson. Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526–557, 1988
8. D.R. Swanson. Complementary structures in disjoint science literatures. In A. Bookstein, Y. Chiamarella, G. Salton, and V.V. Raghavan, editors, *Proceedings of the 14th Annual ACM Conference of Research and Development in Information Retrieval (SIGIR'91)*, pp. 280–289. ACM, New York, 1991

9. D.R. Swanson. Atrial fibrillation in athletes: implicit literature-based connections suggest that overtraining and subsequent inflammation may be a contributory mechanism. *Medical Hypotheses*, 66:1085–1092, 2006
10. J.M. Ziman. *Public Knowledge*. Cambridge University Press, Cambridge, 1968
11. J.M. Ziman. Information, communication, knowledge. *Nature*, 224:318–324, 1969



<http://www.springer.com/978-3-540-68685-9>

Literature-based Discovery

Bruza, P.; Weeber, M. (Eds.)

2008, XII, 198 p. 25 illus., Hardcover

ISBN: 978-3-540-68685-9