Journal of the American Society for
Information Science and Technology

Literature Based Discovery: Beyond the ABCs

Journal:	*Journal of the American Society for Information Science and Technology*
Manuscript ID:	JASIST-2011-04-0161.R1
Wiley - Manuscript type:	Review Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Smalheiser, Neil; University of Illinois at Chicago, Psychiatry
Keywords:	text mining < data mining < knowledge discovery < automatic extracting < computer applications < computer operations < (activities and operations), knowledge discovery < automatic extracting < computer applications < computer operations < (activities and operations)

Literature Based Discovery:

Beyond the ABCs

By

Neil R. Smalheiser

University of Illinois at Chicago
Psychiatric Institute MC912
1601 W. Taylor Street
Chicago, IL 60612
312-413-4581
neils@uic.edu

Abstract

Literature based discovery (LBD) refers to a particular type of text mining that seeks to identify non-trivial assertions that are *implicit*, and not explicitly stated, that are detected by juxtaposing (generally a large body of) documents. In this review, I will provide a brief overview of the past and present of literature based discovery, and will propose some new directions for the next decade. The prevalent A-B-C model is not “wrong”. However, it is only one of several different types of models that can contribute to the development of the next generation of LBD tools. Perhaps the most urgent need is to develop a series of objective literature-based interestingness measures, which can customize the output of LBD systems for different types of scientific investigations.

Introduction

Text mining is an umbrella term for extracting and analyzing information expressed in the form of text. Literature based discovery (LBD) refers to a particular type of text mining that seeks to identify non-trivial assertions that are *implicit*, and not explicitly stated, within (generally a large body of) documents. As articulated by Don Swanson (1986a,b, 1988), identifying such assertions is a first step in formulating and assessing new scientific hypotheses that may be regarded as potential new discoveries. Strategies for literature based discovery have been studied primarily by information and computer scientists (see the comprehensive book edited by Bruza and Weeber (2008) for reviews (e.g., Hristovski et al, 2008; Sehgal et al, 2008; Smalheiser and Torvik, 2008; Wren, 2008; Yetisgen-Yildiz and Pratt, 2008). The bioinformatics community has also created numerous specialized systems that utilize implicit textual assertions for predicting, e.g., gene associations with disease and protein-protein interactions (e.g., Jansen et al, 2003; Rzhetsky et al, 2007; Leach et al, 2009; van Haagen et al, 2009; Tjioe et al, 2010). In this review, I will provide a brief overview of the past and present of literature based discovery, and will propose some new directions for the next decade.

The goal of literature-based discovery is really to generate or assess new **hypotheses** which might represent potential scientific discoveries, and hence are worthy of follow up in the laboratory or clinic. The term literature-based discovery can be ambiguous or misleading (Kostoff, 2007, Kostoff et al, 2009) and Bekhuis (2006) has proposed that it should be replaced with some alternative term such as “exploratory mining”. “Discovery” has many different meanings in different contexts and at different stages in the cycle of scientific discovery (Grinnell, 2009). A LBD system might be very useful when it "discovers" things that are novel to the investigator doing the search, even if it is well known to other experts or even to the scientific community at large. On the other hand, a great deal of information has been published, and hence *ought* to be known by the scientific community, yet lies unknown, unaccessible or neglected for one reason or another ("undiscovered public knowledge"; Swanson, 1986a; "neglected medical discoveries", Swanson, 2011).

A few years ago, Vetle Torvik and I published a case of "undiscovered public knowledge" in genomics databases – namely, the fact that a significant subset of mammalian microRNA precursors derive entirely from genomic repeat elements (Smalheiser and Torvik, 2005). To make this observation, all that was necessary was to view microRNA genes on the UCSC Genome Browser, juxtapose the microRNA track with the Repeatmasker track, and notice the association. The knowledge contained in the Browser is entirely public and explicit; nothing implicit was involved. However, no one had apparently thought to look for such a pattern before -- it was literally hidden in plain view.

This single discovery can be deconstructed into a series of discoveries: First, in the course of an earlier study (Smalheiser and Torvik, 2004), we "discovered" the hypothesis that microRNAs might derive from genomic elements; then, we "discovered" the observation as empirical data lying within public databases. Finally, the finding was analyzed further in detail, written up, and subjected to peer review, to establish the microRNA / genomic repeat link as a generally accepted and biologically significant fact, which would be generally acknowledged as a "discovery" by anyone's definition (Grinnell, 2009). With these caveats, in the present paper, I will refer to any knowledge or finding identified using a LBD system or strategy as a "discovery" regardless of where it sits in the cycle of scientific discovery – as long as it provides something new to the searcher that assists him or her in the task of generating or assessing a hypothesis.

A "Dirty Little Secret"

A further ambiguity is that literature-based discovery can refer either to a "system" – that is, a software product designed to assist (or replace) humans in formulating hypotheses – or to a "strategy" – a cognitive approach that humans employ to combine assertions, whether carried out as a deliberate conscious effort or in an intuitive manner. For several reasons, it has been very difficult to obtain hard evidence documenting the extent to which literature-based discovery does, or potentially can, accelerate the process of scientific discovery.

One the one hand, only a score or so published scientific articles have proposed hypotheses that they said were obtained via literature based discovery systems; only a few have validated the hypotheses experimentally in the same article (e.g., Wren et al, 2004) or even openly acknowledged that LBD played any role in their thinking (Manev and Manev, 2010). Some observers (e.g., Spasser, 1997) have used this paucity of evidence to suggest that LBD arose within the information science community (and stayed there) without successfully connecting with active scientists. However, we must not forget the stark distinction between Private and Public phases of discovery– most of the thoughts, conjectures, pilot studies, puzzling findings, modeling activities, and literature searches that are pursued during the private phase of a scientist's work are missing, sanitized or erased from the final published article (Grinnell, 2009). Just as scientists are generally loath to publish negative findings, most experimental scientists regard hypothesis-papers as an inferior type of literature (in the same manner, I suppose, in which poets regard limericks) and generally will only postulate new hypotheses in print when tacked at the end of an experimental study or a review article. Another factor is that scientists may be reluctant to trust, much less give credit to LBD systems for their outputs. Computer-generated diagnosis systems were rejected by physicians, in part, for similar reasons – they were unwilling to trust or to credit the software when it gave the correct diagnosis, since physicians still had to double-check its reasoning and use their own judgment anyway (Shortliffe, 1987).

More likely, scientists are, indeed, carrying out LBD analyses routinely on their own, manually and unsystematically, perhaps without realizing it. For example, Don Swanson once followed up the impact of several of his classic LBD hypothesis articles (Swanson, 1986b, 1988) by looking at later articles written by others, which validated these hypothesis in experimental or clinical studies. He demonstrated persuasively that these later authors had read, and been influenced by, his own earlier papers, yet few of them cited or discussed them (Swanson, 1993). Moreover, at the panel "Beyond (simple) Reading: Strategies, Discoveries, and Collaborations" held at the 2009 ASIS&T meeting, I gave a detailed example of one neuroscientist who carried out a classic, systematic A –

B – C analysis that led to the discovery of a new extracellular matrix protein receptor – yet was unaware that she was performing a discrete, iterated LBD text mining task. She thought she was simply reading a bunch of articles and reasoning logically about them! Indeed, literature based discovery does represent intuitive common sense, but domain scientists do not realize that modeling common sense is a formal (and very hard) problem.

To my knowledge, there has not been any systematic evaluation of when, and how often, scientists carry out LBD-style analyses (manually) in the course of their scientific work. Nor is it clear whether scientists, themselves, recognize when they are doing a LBD analysis, as opposed to carrying out a literature search or other types of information seeking activities. This is a great PhD thesis topic for someone.

Yet another hurdle for the LBD community is the fact that most domain scientists in the biomedical and physical sciences seem to be unaware of the various web-based LBD interfaces that have been set up by information scientists (reviewed in Weeber et al, 2005). Only a few of these websites have been maintained continuously by their creators, and only a few have been subjected to user testing (Smalheiser et al, 2006; Yetisgen-Yildiz and Pratt, 2008; Yetisgen-Yildiz et al, 2009). The Arrowsmith two node search interface (http://arrowsmith.psych.uic.edu) has been shown to assist field testers materially in assessing their hypotheses (Smalheiser et al, 2006), and has even garnered unsolicited testimonials from outside users of the site (Best of the Web, 2007; Manev and Manev, 2010).

Finally, hypothesis formation is only one of many driving forces for discovery. Someone may have a good hypothesis and not pursue it for a variety of reasons including lack of funding, lack of available analysis tools (Edwards et al, 2011), competing priorities, prevailing biases, and so on. Given all of these considerations, we should not be unduly discouraged that literature based discovery seems to have a low profile among domain scientists. (Bear in mind that most biomedical scientists do not even utilize informatics tools for other basic tasks such as visualizing their data, or summarizing the documents retrieved by a literature search.) Going forward, information scientists can raise its profile not only by improving LBD algorithms, but also by studying the prevalence and role of LBD-like analyses in scientific workflow, and by educating both students and scientists in informatics literacy.

Incremental vs. Radical Discoveries

Swanson formulated the strategy of literature-based discovery in terms of what has become known as the ABC model (Swanson, 1986b, 1988; Swanson and Smalheiser, 1997). For example, given the assertion “A affects B” appearing in one article, and “B affects C” appearing in a different article, one can derive the implicit assertion “A affects C” which represents a potential hypothesis. This formulation has simplicity and power, and (given a corpus of articles of the size of PubMed) suffices to generate an enormous number of plausible hypotheses. Nevertheless, the time has come to relax the ABC formulation and consider alternatives for the field of literature-based discovery.

The ABC approach, as commonly pursued, begins with a collection of articles “A” within MEDLINE or PubMed that represent a scientific problem (e.g., articles that discuss small cell lung carcinoma). Words and phrases “B_i” (which appear in the title or abstract of articles in A) are then listed, and for each “B_i” term (or a filtered subset), a separate literature search is carried out using that term as query. The words and phrases “C_i” which appear in each of the Bi literatures are then compiled (and possibly filtered). Finally, by some criteria, the C_i terms are ranked, such that high ranking C_i terms are thought to represent the most promising hypotheses. (Depending on the system, B_i and C_i may alternatively represent other features extracted from the articles such as Medical Subject Headings or concepts.)

For example, for A = small cell lung carcinoma, and C = members of the category of therapeutic agents, the C_i terms may be the names of drugs which have not yet been tested against small cell lung carcinoma, but which have been proven to have efficacy in other situations (e.g. in other forms of cancer or in animal models) suggesting that they might be explored as new therapies. (Note: some authors reverse the A and C in this scheme, so that one begins with a problem C and seeks to find a possible solution A.)

There are several limitations in this ABC approach. First, the sheer number of B_i terms causes an exponential explosion that is hard to handle computationally, and which requires one or more short-cuts to be implemented (Wren, 2008). Second, the huge number of resulting C_i terms is difficult to assess or interpret manually, so that it is crucial to have effective ranking procedures in order to identify the most promising finds.

Although different systems have dealt with these two issues in various ways, almost all current systems employ **similarity algorithms** that rank C_i terms as more promising if they closely resemble terms or concepts that are already known to be true in A. For example, thalidomide has been investigated as a therapy against certain autoimmune diseases, and a LBD analysis predicted that it may be worth investigating in certain other diseases that share similar pathogenetic features (Weeber et al, 2003). Reelin has been shown to bind to certain proteins, and a LBD analysis identified other proteins (that share certain features with the known set) as promising reelin-binding proteins (Homayouni et al, 2005). By their very nature, similarity algorithms will only find incremental discoveries – those that are similar to what in machine learning is called "the training set" (see also Kostoff et al, 2009).

Another, more subtle limitation of the ABC approach is that systems are generally evaluated according to the probability that the AC_i assertions are likely to be true. That is, they look for highly probable assertions. However, novel discoveries often seem very *unlikely* at the time that they are first proposed (Simonton, 2004). A better approach is to rank the C_i terms according to how many different biological mechanisms link Ci and A, but the sheer number of linking B_i terms (e.g., as tabulated by Don Swanson's Kiwi 1-node search system; Swanson and Smalheiser, 1987) is a poor proxy for estimating this. Other methods, such as mutual information measure, have also been proposed (Wren, 2004). Use of directional action cues (does A inhibit or enhance B? Giles and Wren,

2008) and mapping genes or terms onto functional pathways (e.g., Kim et al, 2011) are active research areas in bioinformatics and may contribute to the solution of this problem.

Moreover, several of the discovery systems attempt to improve the signal-to-noise ratio by employing natural language processing techniques that identify explicit statements of the form “A affects/binds/regulates/interacts with B” and “B affects/binds/regulates/interacts with C” (e.g., Hristovski et al, 2008). This is certainly a valid approach, particularly suited to simple statements of chemical interactions, and useful for genomics and proteomics data in particular.

However, I argue that most implicit information present in the scientific literature does not follow such simple templates (and may not consist of simple factual or propositional statements at all). Rather, it is analogies and images -- juxtapositions and novel associations of ideas – that appear most often to stimulate scientists to formulate radically new hypotheses (see discussion in Simonton, 2004). Many classic discoveries follow AB and BC assertions but at a rather high level of abstraction that is unlikely to be captured or highlighted in explicit templated factual statements:

a) According to Lenoir and Giannella (2006): “The technological development of peptide and DNA microarrays was driven by analogy to photolithography techniques, particularly those employed by the semiconductor industry. In one of the meetings of the Affymax scientific board, Leighton Read tossed out the idea of just mimicking the makers of semiconductor chips, who use beams of light to manipulate molecules on solid surfaces in order to create random chemical diversity”.

b) According to Ban (2006): “Potassium bromide is the oldest widely used sedative in medicine. Charles Lockock, a London internist, discovered the anticonvulsant and sedative action of the drug. His discovery was one of the many quaint examples of serendipity in which an utterly false theory led to correct empirical results. Lockock, like most physicians of his time, believed that there was a cause-effect relationship between masturbation, convulsions, and epilepsy. Bromides were known to curb the sex drive. Lockock’s rationale was to control epilepsy, i.e., convulsions, by reducing the frequency of masturbation. The treatment was a success insofar as control of convulsions was concerned. It also brought to attention the sedating properties of the drug.” (Admittedly one could construct this discovery from individual pre-existing statements, but only if one were to accept *false* statements (thought to be true at the time) as inputs for discovery systems!)

c) In my own scientific work, we proposed that RNA interference may have a physiologic role in regulating learning and memory (Smalheiser et al, 2001). This hypothesis was based on similarities between gene silencing studies in *C. elegans* that were published around 2000, and experiments carried out on memory transfer in planarians more than 30 years earlier. For example, 1) one can feed *C. elegans* bacteria that express double-stranded RNAs to induce silencing; whereas one could transfer memory in planarians by feeding naïve worms extracts of trained worms. 2) One can inject double-stranded RNAs in one location and it will spread gene silencing throughout the body of *C. elegans*; whereas one could cut off the foot of a trained planarian and it would regenerate a new head that retains the memory. 3) The silencing activity in *C. elegans* depends on double-stranded RNAs; whereas the active memory transfer molecules in planarians appeared to be some type of RNA. 4) RNA interference in *C. elegans* is extremely potent and self-amplifying; whereas memory transfer in planarians was effective even when the extracts did not contain any detectable RNA at all (at levels that were measurable by optical density).

Even if each of these individual similarities could be captured in simple templated factual assertions within a body of articles within each literature (which is doubtful, at least for the primary research articles), no single feature was very compelling, specific, or unusual, so it is unlikely that they would have drawn attention in the forward direction from a discovery system. Rather, it was the combination of all four similarities that created an intriguing story and led to the testable hypothesis that endogenous siRNAs are expressed in brain and up-regulated during the onset of learning (Smalheiser et al, 2001). Interestingly, the initial experimental attempts to detect endogenous siRNAs (during 2000-2005) all gave negative results. This did not disconfirm the hypothesis, however, since the recent development of deep sequencing methodology has allowed them to be reliably detected (see discussion in Smalheiser et al, 2011).

Another limitation of the NLP-based approach, i.e., utilizing templated assertions, is that they often enforce semantic agreement across the linking term. That is, to link AB and BC assertions, the term B must have the same meaning or context in both AB and BC. Yet Magnesium itself can be mapped to many different concepts – it can be conceptualized as an element, a cation, a dietary ingredient, a bodily fluid constituent, a co-factor of enzymes, a channel blocker, or a therapeutic agent. The same term (Mg) is often discussed in different contexts in different literatures that we would like to connect. The limited "slippage" across those loose links is *desirable*, and may be lost if links are forced to share the same semantic meaning or connotation. Root-Bernstein (1989) gave an example of the importance of slippage in the discovery of lysozyme by Alexander Fleming: "Enter Fleming the mischievous game player. His problem: What causes his frequent and uncomfortable runny noses? Wait a minute! Runny bottoms are caused by bacteriophage infections! Why not runny noses? A hypothesis is born of verbal analogy!"

Interestingness Measures for Literature Based Discovery Systems

To date, the challenge of literature-based discovery (the one node search) has largely been framed in terms of finding hypotheses that are **novel**, **non-trivial**, and **likely to be true**. On the other hand, Torvik and Smalheiser (2007) employed shared title words and phrases (B-terms) to link two disparate literatures A and C in a biologically meaningful manner, in which the emphasis was on finding terms that are **relevant** and **meaningful** in a particular context**.** Yet, significant scientific discoveries have one or more additional aspects: For example, they may exhibit **simplicity**, they may be **surprising**, or **beautiful** in an aesthetic or conceptual sense. They often link disparate **disciplines**, and ideally they are **actionable** (i.e., they lead to testable hypotheses that can be tested immediately or in the near future). They have great **impact** within their own field, their premises are based on reliable experimental **support**, and they have **explanatory power** that generalizes and ripples widely across other domains of science.

Whereas the field of numerical data mining has extensively explored a variety of rule interestingness measures (Han and Kamber, 2006), to my knowledge, few interestingness measures have been formulated in the context of text mining, and even fewer have applied literature-based measures (e.g., Weiss et al, 2010; Sebastian and Then, 2011). Interestingness measures can be objectively formulated for a given finding “A affects B” in terms of formulas that are derived from literature based features (i.e., the set of articles that demonstrate, mention or discuss “A affects B”) or literature pairs (i.e., the set of articles related to A and the set of articles related to B). The study of Swanson et al. (2001) was a case in which interestingness measures were employed to identify viruses that were particularly promising to be exploited for biological warfare. The premise was that bio-warfare investigators were most likely to choose viruses which had their genomes already sequenced and which had been investigated with regard to aerosol stability. (This strategy is based on a model of how biological warfare researchers may themselves select a virus for study.) Thus, the list of potential viruses was ranked according to how **actionable** they were for experimental manipulation. A parallel study by Smalheiser (2001) used similar criteria to predict that gene therapy biotechnologies (specifically, gene delivery methods) were likely to be employed for viral bio-warfare research.

Removing the "B" from the A-B-C Model: Reformulating One-Node Searches as Two-Node Searches

As mentioned above, one-node searches have generally been formulated in a manner that faces an explosion of intermediate links: Starting with a single literature A, one obtains up to thousands of B_i-terms, and for each B_i-term, a new query is performed that obtains many C_i-literatures. Because of this, all existing LBD strategies restrict the number or type of B-terms, and most restrict the C_i-literatures to those that fall within a pre-determined category (e.g., diseases or drugs). Yet one can bypass the process of collecting B-terms altogether, at least for the purpose of identifying candidate C_i-literatures (Torvik and Smalheiser, 2007). This is because the range of possible C_i-literatures are generally known in advance. Given a specific disease (say, A = Parkinson's disease), we may be looking for novel therapeutic agents – say, the C_i-literatures may comprise the list of drugs that are FDA-approved for *other* indications but not previously tested in Parkinson's disease. One simply makes a list of **all** agents within the general category, and examines them one by one. In other words, a one-node search can be performed by carrying out a series of two-node A-C_i searches, in which the output from each search is a score that estimates how good C_i is as a candidate. One simple score is the estimated amount of overall shared implicit information that is shared between the A and C_i-literature (Torvik and Smalheiser, 2007), though it is likely that better rankings will be achieved using a combination of interestingness measures. Certainly, the B_i-terms are not irrelevant to this process, since they are likely to be useful features in calculating the overall scores for each two-node search. Yet, they no longer sit as a bottleneck in the discovery system.

A Phone Call from Don

My first contact with Don Swanson occurred in the early 1990's, when he phoned me to discuss an apparent anomaly in his analyses. Following up on his Mg-migraine hypothesis (Swanson, 1988), he had noticed that Mg seemed to rank highly as a candidate therapy, no matter what neurological disease was under consideration. How could this happen? I said the issue was very simple: Mg is known to gate (i.e., limit) calcium currents through the NMDA receptor. Over-stimulation of the NMDA receptor, or over-accumulation of intracellular calcium, causes excitotoxicity, which occurs in many diverse situations (stroke, ALS, seizures, etc.). Thus, a deficiency of Mg should exacerbate excitotoxicity and Mg supplementation should help to counteract it, not just in migraine, but across many neurological diseases. In fact, our first joint paper pointed this out in the context of individuals who exhibit mild dietary Mg deficiency (Smalheiser and Swanson, 1994). Putting this back in terms of the A-B-C model, one could say that the candidate C_i = Mg is highly **interesting** regardless of the specific A literature, at least within a certain range.

Whereas measures to identify emerging research fronts have been the concern of scientometrics and bibliometrics, these measures have tended to be geared towards policy makers and sociologists -- detecting the fronts after they have already started to become “hot”. Some areas are not simply “hot”, but have such pervasive implications (noncoding RNAs, prion proteins, microRNAs) that they should arguably be ranked high on any list of possible topics to study, no matter what the specific question and regardless of the specific area of interest by the investigator. This is reminiscent of a t-shirt slogan that I have seen: “No matter what the question is… the answer is to do more yoga.”

Nevertheless, most scientists are likely to feel that they can identify “hot” areas already. The biggest need, and the biggest “bang for the buck” for literature-based discovery, is to identify research areas that are currently *neglected*, but which, when juxtaposed with other information, have the potential to identify important frontier areas for investigation (Smalheiser and Torvik, 2008; Swanson, 2011). There are many reasons why a line of work may have become neglected; these need not be discussed here. However, one would like to reconsider and possibly revive those neglected hypotheses or lines of work that are the most **interesting** when viewed in light of other more recent evidence that have appeared in other scientific fields -- even if – perhaps especially if -- the original hypotheses were generally thought to be “wrong” or experimentally disproved.

Inheritance of acquired characteristics is a stellar example of a field that, for more than a hundred years, appeared to be a pre-Darwinian relic that was thoroughly discredited as scientific nonsense. Recent findings in genomics and molecular biology, however, have validated several mechanisms by which environmental stimuli can influence the genome and pass changes to subsequent generations (Landman, 1991; Liu, 2007; Koonin and Wolf, 2009). In fact, this area has quickly become one of the "hottest" in biomedical science. The studies on memory transfer in planarians (discussed above) is another example of a field that was abandoned after the original practitioners had retired, yet sparked a new field of investigation.

Once again, Don Swanson has pioneered the effort to identify neglected research findings, which he conceptualized as a generalization of one-node searching (Swanson, 2011). However, much more work is needed to discern which neglected findings ought to remain that way; which deserve revival; and which (when combined with other findings) create an entirely new and promising hypothesis.

The Problem of Creating Gold Standards for LBD Systems

In order to evaluate and compare different LBD systems, it is crucial to develop an extensive set of gold standard examples. The very nature of one-node searches and their traditional goal (to identify totally novel hypotheses with no existing experimental support) makes it difficult to establish gold standards (Smalheiser and Torvik, 2008). Some studies have employed a handful of validated one-node searches created by Swanson's early predictions (Swanson, 1986b, 1988) and others have advocated the use of time-sliced literatures to evaluate LBD methods. In this approach, LBD predictions are based upon an analysis of MEDLINE at a given date. One examines MEDLINE articles at later dates to see if the predictions have been confirmed or at least investigated subsequently. Another option is to employ lists of known facts or relationships, either extracted from the literature or manually curated, as an external standard for one-node searches (e.g., Homayouni et al, 2005). For example, suppose one is conducting a LBD analysis to predict novel interactions that reelin may have with other proteins. Given a list of proteins known to interact with reelin, a successful LBD method should rank the known interactors highly, even if they are excluded from the final list of predictions due to lack of novelty.

Besides these evaluation methods, one can imagine innovative ways of utilizing other datasets. For example, the TREC Genomics 2006 and 2007 queries resemble one-node searches insofar as they seek to rank articles within a given category (equivalent to the Ci-literatures) in terms of their relevance to a given item or concept (equivalent to literature A). Thus, if one were to apply one-node search systems to these data, one could employ the gold standard TREC results. Another idea is to obtain the abstracts of new R01 and R21 grants that have been funded by NIH, available via the CRISP/RePORTER database. Certainly, at the time the grant was reviewed, a panel of experts had agreed that the central aims were novel and promising for further study – so a good LBD system should be able to identify them and rank them highly. Similarly, new hypotheses that are proposed in a published review article can be regarded as a gold standard of what (at least certain) experts feel are promising new research directions. The search for different ranking strategies and the project to build gold standards should proceed in parallel, covering a variety of different ranking strategies, since a strategy to identify relevant information will be expected to rank items quite differently than one intended to identify high-risk, paradigm-shifting ideas.

Concluding Thought

The A-B-C model is not “wrong”. However, it is only one of several different types of models that can contribute to the development of the next generation of LBD tools. Perhaps the most urgent need is to develop a series of objective literature-based interestingness measures, which can customize the output of LBD systems for different types of scientific investigations. The field of bioinformatics has exploded in the past few years, due to the richness of genomics and proteomics datasets, despite employing (for the most part) relatively simple data mining, statistics and text-based mining methods. The scientific literature is certainly rich enough, and expanding rapidly enough, for literature-based discovery systems to serve as major facilitators of scientific discovery.

Acknowledgements

This paper is an expanded version of a presentation made at the 2009 ASIS&T Annual Meeting. It is dedicated to my senior partner, Don Swanson, and my junior partner, Vetle Torvik, who helped shape the ideas in this article and arguably should have been co-authors (except that then I could not make this dedication to them!).

References

Ban, T.A. (2004). The role of serendipity in drug discovery. Dialogues in Clinical Neuroscience, 8, 335-344.

Bekhuis, T. (2006). Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. Biomedical Digital Libraries, 3, 2.

Bruza, P. & Weeber, M. (Eds.) (2008). Literature-based discovery. Berlin: Springer-Verlag.

Best of the Web. (2007). Genetic Engineering & Biotechnology News, 27, 20.

Edwards, A.M., Isserlin, R., Bader, G.D., Frye, SV., Willson, T.M., & Yu, F.H. (2011). Too many roads not taken. Nature, 470, 163-165.

Giles, C. B. & Wren, J. D. (2008). Large-scale directional relationship extraction and resolution. BMC Bioinformatics, 9 Suppl 9, S11.

Grinnell, F. (2009). Everyday practice of science. New York: Oxford University Press.

Han, J. & Kamber, M. (2006). Data mining: Concepts and techniques, 2nd ed. New York: Elsevier.

Homayouni, R., Heinrich, K., Wei, L., & Berry, M.W. (2005). Gene clustering by latent semantic indexing of MEDLINE abstracts. Bioinformatics, 21, 104-115.

Hristovski, D., Friedman, C., Rindflesch, T.C., & Peterlin, B. (2008). Literature-Based Knowledge Discovery using Natural Language Processing. In Bruza, P. & Weeber, M. (Eds.) Literature-based discovery (pp.133-152). Berlin: Springer-Verlag.

Kim, Y.-A., Wuchty, S., & Przytycka, T.M. (2011). Identifying Causal Genes and Dysregulated Pathways in Complex Diseases. PLoS Computational Biology, 7, e1001095.

Kostoff, R.N. (2007).Validating discovery in literature-based discovery. Journal of Biomedical Informatics, 40, 448-450.

Kostoff, R. N., Block, J. A., Solka, J. L., Briggs, M. B., Rushenberg, R. L., Stump, J. A., Johnson, D., Lyons, T. J., & Wyatt, J. R. (2009). Literature-related discovery. Annual Review of Information Science and Technology, 43, 1–71.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., & Gerstein M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science, 302, 449-453.

Koonin, E.V. & Wolf, Y.I. (2009). Is evolution Darwinian or/and Lamarckian? Biology Direct 4, 42.

Landman, O.E. (1991). The inheritance of acquired characteristics. Annual Review of Genetics 25, 1-20.

Leach, S.M., Tipney, H., Feng, W., Baumgartner, W.A., Kasliwal, P., Schuyler, R.P., Williams, T., Spritz, R.A., & Hunter, L. (2009). Biomedical discovery acceleration, with applications to craniofacial development. PLoS Computational Biology, 5, e1000215.

Lenoir, T., & Giannella, E. (2006). The emergence and diffusion of DNA microarray technology. Journal of Biomedical Discovery and Collaboration, 1, 11.

Liu, Y. (2007). Like father like son. A fresh review of the inheritance of acquired characteristics. EMBO Reports 8, 798-803.

Manev, H., & Manev, R. (2010). Benefits of neuropsychiatric phenomics: example of the 5-lipoxygenase-leptin-Alzheimer connection. Cardiovascular Psychiatry and Neurology, 2010, 838164.

Root-Bernstein, R.S. (1989). How scientists really think. Perspectives in Biology and Medicine, 32, 472-488.

Rzhetsky, A., Wajngurt, D., Park, N., & Zheng, T. (2007). Probing genetic overlap among complex human phenotypes. Proceedings of the National Academy of Sciences USA, 104, 11694-11699.

Sebastian, Y. & Then, P.H.H. (2011). Domain-driven KDD for mining functionally novel rules and linking disjoint medical hypotheses. Knowledge-based Systems, 24, 609-620.

Sehgal, A.K., Qiu, X.Y. & Srinivasan, P. (2008). Analyzing LBD methods using a general framework. In Bruza, P. & Weeber, M. (Eds.) Literature-based discovery (pp. 75-100). Berlin: Springer-Verlag.

Shortliffe, E.H. (1987). Computer programs to support clinical decision making. Journal of the American Medical Association, 258, 61-66.

Simonton, D.K. (2004). Creativity in Science: Chance, logic, genius, and Zeitgeist. Cambridge, UK: Cambridge University Press.

Smalheiser, N.R. (2001). Predicting emerging technologies with the aid of text-based data mining: a micro approach. Technovation, 21, 689-693.

Smalheiser, N.R., Lugli, G., Thimmapuram, J., Cook, E.H., & Larson, J. (2011). Endogenous siRNAs and noncoding RNA-derived small RNAs are expressed in adult mouse hippocampus and are up-regulated in olfactory discrimination training RNA, 17, 166-181.

Smalheiser, N.R., Manev, H., & Costa, E. (2001). RNAi and brain function: was McConnell on the right track? Trends in Neurosciences, 24, 216-218.

Smalheiser, N.R, & Swanson, D.R. (1994). Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. Neuroscience Research Communications, 15, 1-9.

Smalheiser, N.R., & Torvik, V.I. (2004). A population-based statistical approach identifies parameters characteristic of human microRNA-mRNA interactions. BMC Bioinformatics, 5, 139.

Smalheiser, N.R., & Torvik, V.I. (2005). Mammalian microRNAs derived from genomic repeats. Trends in Genetics, 21, 322-326.

Smalheiser, N.R., & Torvik, V.I. (2008). The place of literature-based discovery in contemporary scientific practice. In Bruza, P. & Weeber, M. (Eds.) Literature-based discovery (pp. 13-22). Berlin: Springer-Verlag.

Smalheiser, N.R., Torvik, V.I., Bischoff-Grethe, A., Burhans, L.B., Gabriel, M., Homayouni, R., Kashef, A., Martone, ME., Perkins, G.A., Price, D.L., Talk, A.C., & West, R. (2006). Collaborative development of the Arrowsmith two node search interface designed for laboratory investigators. Journal of Biomedical Discovery and Collaboration, 1, 8.

Spasser, M.A. (1997). The enacted fate of undiscovered public knowledge. Journal of the American Society for Information Science, 48, 707-717.

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

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

Swanson, D.R. (1988). Migraine and magnesium: eleven neglected connections. Perspectives in Biology and Medicine, 31, 526-557.

Swanson, D.R. (1993). Intervening in the life cycles of scientific knowledge. Library Trends, 41, 606-631.

Swanson, D. R. (2011). Literature-based resurrection of neglected medical discoveries. Journal of Biomedical Discovery and Collaboration, 6, 34-47.

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

Swanson, D.R., Smalheiser, N.R., & Bookstein, A. (2001). Information discovery from complementary literatures: categorizing viruses as potential weapons. Journal of the American Society for Information Science and Technology, 52, 797-812.

Tjioe, E., Berry, M.W., & Homayouni, R. (2010). Discovering gene functional relationships using FAUN (Feature Annotation Using Nonnegative matrix factorization). BMC Bioinformatics, 11 Suppl 6, S14.

Torvik, V.I., & Smalheiser, N.R. (2007). A quantitative model for linking two disparate sets of articles in Medline. Bioinformatics, 23, 1658-1665.

van Haagen, H.H., 't Hoen, P.A., Botelho Bovo, A., de Morrée, A., van Mulligen, E.M., Chichester, C., Kors, J.A., den Dunnen, J.T., van Ommen, G.J., van der Maarel, S.M., Kern, V.M., Mons, B., & Schuemie, M.J. (2009). Novel protein-protein interactions inferred from literature context. PLoS One, 4, e7894.

Weeber, M., Vos, R., Klein, H., De Jong-Van Den Berg, L.T., Aronson, A.R., & Molema, G. (2003). Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. Journal of the American Medical Informatics Association, 10, 252-259.

Weeber, M., Kors, J.A., & Mons, B. Online tools to support literature-based discovery in the life sciences. Briefings in Bioinformatics, 6, 277-286.

Weiss, S. M., Indurkhya, N., & Apte, C. V. (2010) Predictive Rule Discovery from Electronic Health Records. Proceedings of the 1st ACM International Health Informatics Symposium *IHI'10,* November 11–12, 2010, Arlington, Virginia, USA, pp. 734-743.

Wren, J. D. (2004). Extending the mutual information measure to rank inferred literature relationships. BMC Bioinformatics, 5, 145.

Wren, J.D. (2008). The 'open discovery' challenge. In Bruza, P. & Weeber, M. (Eds.) Literature-based discovery (pp. 39-55). Berlin: Springer-Verlag.

Wren, J.D., Bekeredjian, R., Stewart, J.A., Shohet, R.V., & Garner, H.R. (2004). Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics, 20, 389-398.

Yetisgen-Yildiz, M., & Pratt, W. (2008). Evaluation of literature-based discovery systems. In Bruza, P. & Weeber, M. (Eds.) Literature-based discovery (pp. 101-113). Berlin: Springer-Verlag.

Yetisgen-Yildiz, M., & Pratt, W. (2009). A new evaluation methodology for literature-based discovery systems. Journal of Biomedical Informatics, 42, 633-643.