Book Review

Literature-based Discovery. Edited by Peter Bruza and Marc Weeber. Berlin: Springer, 2008. xii, 198 pp. $119.00 (hardcover). ISBN: 978-3-540-68685-9. Series: Information Science and Knowledge Management, Vol. 15.

Literature-based Discovery is an interesting and enlightening compilation of 11 articles on literature-based discovery (LBD). The topic is very current, debatable, and I believe fairly original in today's information sciences. Although we can observe with the editors that LBD is not owned by any particular discipline that could claim ownership (e.g., knowledge discovery, text mining, or information retrieval [IR]), it is clearly of foundational interest for several fields of information science. The articles are written from different perspectives and reflect different shades of opinion, with a focus on the pioneering works by Swanson. References and notes of various lengths are included in each article. Although there are several single articles written on LBD, this is, to our knowledge, the first book-length treatment of the topic.

Defining the readership targeted by the book is more challenging. While LBD researchers originate from disciplines as various as machine learning, information science, information retrieval, logics, bioinformatics, and the biomedical sciences, the authors' intended objective is mainly to "inspire new researchers" by establishing the foundation of LBD, thus suggesting that the book is more appropriate for master students and young researchers.

The book is conventionally divided between a general introduction (General Outlook and Possibilities) and a section dedicated to Methodologies and Applications. It is thus expected that the reading complexity should grow gradually until the presentation of a set of successful applications, which should be more readable for nonspecialists. The expectation is somehow wrong, as several chapters in each section could have been understood without prior specific knowledge on LBD and related fields, whereas some others, including in the application section, do demand a significant effort for newcomers.

The first article, "Literature-Based Discovery in Contemporary Scientific Practice" provides a very readable introduction on various key concepts of LBD. In particular, it presents the reader with the model of *complementarity*, which is central for LBD. Also called the ABC model, such a model can be described as follows:

1. Given a relationship between entity A and entity B in article P1;
2. Given a relationship between entity B and entity C in article P2;
3. One can hypothesize that the relationship applies also to A and C.

Such a complementarity principle supposes the availability of some common background knowledge, including a shared language, which is expected in a given scientific community. The introduction also assumes a methodological separation between what can be—often implicitly—known by a particular researcher and what has been published somewhere, thus constituting the *recorded* knowledge, i.e., a sort of official or legacy knowledge. The model also assumes that there exists only a limited amount of connections between the complementary articles. This condition serves to separate the well-known/trivial truth vs. novelty, which supports the notion of *complementary but disjoint* (CBD) literatures. Being able to identify CBD literatures is the key issue in LBD.

This chapter is a synthetic, well-documented, and successful vulgarization effort to introduce the reader to LBD. The chapter supports an implicit theory of science that conforms to a human-centric vision of modern science; although, to some extent LBD could potentially question such a classical vision. Indeed, in this article LBD is seen as a way to enhance human ability—readability?—to identify complementary papers. The author mainly describes the development of user-based LBD; therefore, the impression that LBD could be reduced to a query-driven information retrieval task is suggested.

Formally, the next paper is different from the previous regarding the absence of an abstract and the availability of keywords. The previous paper had indeed no keywords but had a nice abstract! The paper nicely presents concrete examples, including screenshots, showing how specific text mining softwares can be used to guide LBD. Basically, existing systems provide a list of terms, so-called *B-terms*, common to the disjoint scientific domain. The first impression regarding these systems and their effectiveness is disappointing. Indeed, success in LBD seems to refer to metrics such as publication in the peer-reviewed literature, while the main advantage of such systems seems that they are extremely labor-cheap—pretty much like existing MedLine search engines! The rest of the paper explores the difficulty of benchmarking LBD tasks, introducing a strict separation between the literature, analyzed by LBD, and the structure of nature, as analyzed by natural sciences, without apparently noticing that such a classical separation between culture and nature might be again questioned considering that languages are maybe no less natural than biological entities. When comparing TREC-based evaluation and two-node LBD evaluation, the author could have used categories such as *intrinsic* and *extrinsic* evaluation models, i.e., direct evaluation of a task (counting relevant B-Term) versus evaluation of a processing step in some other task (using IR to perform LBD tasks or using LBD to perform IR tasks).

The third article is entitled *The Tip of the Iceberg: The Quest for Innovation at the Base of the Pyramid*. The (lengthy) introduction to the notion of "disruptive" technologies attempts to move the discussion away from the narrow biomedical domain, where LBD seems to have restricted itself to other technological areas. More radically, the authors attempt to move away from scientific discovery to the broader scope of innovation, which does not necessarily require technological advances. The authors consistently argue that literature sources should go beyond specialized scientific libraries, such as MedLine. Indeed it is to be observed that most studies in the field have been applied to the sole MedLine library. By observing that new drugs are often the results of reported drugs' side effects, the authors proposed to invert the traditional discovery workflow assumed from Problems/Hypothesis to Discovery. The demonstration is interesting; however, it has already been recognized that drug discovery is in practice a far more unpredictable process than assumed here (see, e.g., Taleb, 2007). Although difficult to follow, it contains a wealth of ideas, like the invitation to work not only with literature content. Although not explicit in their report, the authors could have relevantly pointed out that the Web is not usually considered by LBD experimentalists. Some other claims could have deserved some development. For instance, when the authors claim that LBD is somehow frequency-independent, so that rare events can be appropriately handled by LBD methods.

In the next article, *The Open Discovery Challenge*, the author proposes to define LBD as a constructive inference process to relate together a string of facts, of which particular links can be well known. Such a synthetic definition is welcome and the article could have come earlier, as it introduces the reader to the rather recent scientific challenge of handling overwhelming quantities of data, including textual or *omics* data and their combination, so-called bibliomics. It explains that computer-driven hypothesis generation is legitimately regarded as a promising but somehow unreachable objective. Thus, confirming the general position defended earlier in the book that LBD is a human-driven interactive task. The article introduces information theoretic measures, such as mutual information, and observes that existing applications (e.g., PubGene, MINT, etc.) do not provide much more than simple associations such as co-occurrences or co-citations. It explains that current literature discovery processes are human labor-intensive and so remain time-consuming. The authors mention more advanced Natural Language Processing methods such as those explored to maintain the MINT database, but unfortunately without further elaborating on what is probably the present and future of LBD (Krallinger, Valencia, & Hirschman, 2008). Interestingly, they also mention the importance of curated data content stored in molecular biology knowledge bases, but they seem to miss the fact that beyond MedLine such data sources can also be used as material for text-based discovery (Mottaz, Yum, Ruch, & Veuthey, 2008). In the same vein, the limitation provided by MedLine abstracts require considering applying LBD to full-text articles rather than abstracts (Natarajan et al., 2006), which clearly suggests that specific regulations, such as those explored by PubMed Central are internationally needed to provide free access to full-text content. The chapter is well written, although it tends to repeat already introduced concepts of LBD. At mid-book, we can regret that the discussion remains very general and fairly historical. In particular, the references—where some are actually duplicated in the article bibliography—are often outdated in a field where trends can radically change in a few years.

R.N. Kostoff, from MITRE, authored the fifth article, *Where Is the Discovery in Literature-Based Discovery*. The author proposes an epistemological investigation on the concepts of discovery and innovation. For the author, *discovery* borrowed its clear definition from artificial intelligence, while *innovation* seems less clearly defined, so that both concepts tend finally to refer to the very same idea in the author's discourse. This is questionable, since innovation, which relates directly to some commercial applications, can proceed without much scientific discovery. In contrast, several scientific discoveries do not translate immediately into marketed products. The rest of the article is dedicated to the validation of discoveries by LBD with important observations such as (1) most efforts of LBD have been carried on in discovery rather than on the validation of these discoveries; (2) for discoveries, most efforts focus on helping early discovery steps (e.g., discovering candidate identification); (3) several reported LBDs were not discoveries! The effort of the author is particularly appreciated considering that the answer to the question he raises might not be to the advantage of LBD. Ultimately, it is suggested that a complete validation process should end up with at least one of the following items: (1) mention of the invention in patents; (2) mention of the discovery in clinical trials; (3) integration of the discovery in medical practice, for instance, evidenced by mention of the discovery in clinical guidelines. While item 1 would mostly apply to innovations, items 2 and 3 would concern scientific discoveries, since item 3 depends normally on the outcome of item 2.

While a differential definition or comparative analysis of LBD versus related fields such as IR or text mining is yet to be provided, the authors of article 6 (*Analyzing LBD Methods Using a General Framework*) clearly decided to adopt a simpler position: knowledge discovery from text, text mining, and LBD are seen as strict synonyms. The synonymy is both arguable and questionable, considering that text mining can be applied to nonpeer-reviewed content (Web, patents, clinical records, etc.), while *Literature*-Based Discovery is supposed to be more specific regarding formatting standards and quality checking. The authors propose a typology to classify LBD systems. The delivered description axes are effective to cover systems presented in the chapter; however, we observe that, like elsewhere in the book, the authors arbitrarily limit their survey to systems based on simple co-occurrences. Consistent with the other contribution, they ignore today's most promising works on Natural Language Processing fact extraction or Question-Answering. As acknowledged by the authors, the resulting report looks more like a partial review on Gene-Disease association systems than as a general typology to broadly classify LBD systems, which is by no mean a trivial task.

Article 7 (*Evaluation of Literature-Based Discovery Systems*) is dedicated to the evaluation of LBD systems; again, systems/experiments presented elsewhere in the book are discussed in this chapter. Although such a redundancy hinders the reading of the full book, it is probably the right approach since readers of such collected papers might not be interested in all aspects of LBD. The authors distinguish four types of evaluations: replicating Swanson's experiments, using classification or retrieval metrics, asking experts to judge the results, and publishing in the medical domain. The former relates to the "impact analysis" already discussed in Kostoff's contribution (e.g., measuring impact by counting the number of patented discoveries), and proposes to evaluate the effectiveness of LBD studies by looking at the journal's impact factor in publishing the study. In practice, such a "validation" method is apparently rarely used, since high-impact journals are usually medical journals, which requires coauthoring with medical researchers!

Article 8 (*Factor Analytic Approach To Transitive Text Mining Using Medline Descriptors*) explores in detail a particular statistical instrument: factor analysis. At first, in particular when reading the abstract, this chapter is a surprise, since the authors' intentions seem unclear, but the introduction is very didactic, with the recall of the seminal experiences pioneered by Swanson (e.g., Raynaud's disease and fish oil, etc.). The author describes the advantages of linear least square fit mapping methods to perform LBD tasks. He argues that such a method can help find quick and easy screening terms to discover interesting associations between various replicated case studies. Those familiar with pseudo-relevance feedback as used in IR can immediately appreciate how latent semantic indexing and the like are appropriate to establish implicit (so-called latent) associations between terms. However, a fair presentation should also state how computationally expensive such approaches are, in particular compared to well-known alternatives such as Rocchio relevance feedback.

Article 9, *Literature-Based Knowledge Discovery Using Natural Language Processing (NLP)*, brings some updated fresh air into the collection; indeed—at last—the authors attempt to go beyond the usual feature (e.g., words, MeSH, genes, etc.) associations. They relevantly observe that computational linguistics approaches are faced with scalability issues, which make them less operational in practice than expected. The authors conclude that hybrid approaches likely to use both fine-grained NLP and co-occurences are promising. Implicitly, they criticize the inconsistency of NLP tools such as MedLEE and SempRep regarding the fact that the two tools tend not to provide compatible results. A more argumentative and synthetic comparison between the power and limits of NLP versus co-occurrences methods could have helped understand the stakes of the discussion as well as the various contexts where NLP can be beneficial and where it is probably not.

Given the importance of IR tasks for LBD, the tenth article (*Information Retrieval in Literature-based Discovery*) is more than welcome. We only regret it was not provided earlier. Not only would it have provided powerful theoretical and typological instruments to define and evaluate LBD, but it could have helped the newcomers to better understand the field by capitalizing on everyone's familiarity with Web and library search engines. As for the theoretical background, a clarification effort could have helped characterized LBD. Thus, the following could have been drafted: LBD is an interactive search and association task applied to digital libraries, often powered with automatic text categorization tools, using document features such as phrases and terms in terminologies and ontologies (e.g., MeSH, Gene Ontology, etc.) and nontextual features (e.g., bibliographical citations). The task can benefit from methods originally designed for automatic or pseudo-relevance feedback such as Rocchio or approaches based on factor analysis. The author proposes

DOI: 10.1002/asi

to define IR as the field concerned with the "indexing and retrieval of knowledge-based information." Such a definition would clearly offer some interesting debates in the information sciences forum (see, e.g., Capurro & Hjorland, 2003). More consensual definitions could have been proposed when defining the purpose of IR such as "the purpose of an information retrieval system was to provide information about a request" (van Rijsbergen & Lalmas, 1996). From a functional perspective and considering information retrieval as a task, indexing should be regarded as no more than a convenient option to achieve good retrieval performance, i.e., efficiency and effectiveness; in particular, Grep-like tools are examples of index-free IR engines. The task to introduce IR is challenging and articulating IR with LBD is by no mean trivial; therefore, more room could have been given to comparison of the two tasks. The "suggested readings" section is particularly welcome for student readers and we regret that such a section is not supplied in each chapter. Together with the list of books provided here, a short list of highly synthetic introduction papers could have been useful, such as the excellent 7-page paper by Singhal (2001), also useful to understand the next article of the book.

The title of the last paper (*Biomedical Application of Knowledge Discovery*) is misleading for two reasons: first, almost the whole book is mostly focused on "Biomedical Application of Knowledge Discovery"; second, this chapter in fact describes a particular LBD system. The content is fairly technical. The discussion on the importance of the document length parameter, which plays a crucial role in relevance-weighting schema used by retrieval engines, would have needed some background introduction in the previous paper. The cited papers (Singhal et al., 1996; Fujita, 2004) are neither more synthetic nor more readable than Singhal's review (2001). Furthermore, the length normalization factor proposed (Lnu) is only one out of a family of pivoted normalization schema (e.g., dtu.dtn), which tends to also perform quite well on MedLine (Aronson et al., 2005) but no more than other schema, such as probabilistic schema (Abdou & Savoy, 2008). In contrast with other papers of the book, the richness and density of this paper is of interest, including for researchers familiar with the subject. Some approximate claims could have been corrected at reviewing time: thus, pharmacogenomics goes beyond single nucleotide polymorphism (SNP) analysis and it is only because current screening methods can hardly analyze polymorphisms beyond single nucleotide deletion/shift that SNPs are of particular interest.

Disregarding the content, and focusing on the editorial make-up of the book, the work suffers from weaknesses regarding its editorial content: the thematic index is disappointing for a book written by text processing experts, references are repeated several times in a given bibliography, while figures and tables sometimes provide a surprisingly poor content (see, e.g., figure 1, p. 78). English grammatical errors are not rare—even for a non-native speaker (e.g., p. 148, "evaluation" is used instead of "evaluated"). Misspellings can also be found in the index: BIOTLA refers to the BITOLA system. Further, there are several implicit and fruitful connections between the various articles of the compilation, and we can regret that the authors have not made the effort to turn them into explicit references. Given the poorness of the index table, such explicit cross-references would have been of primary interest for students.

In conclusion, it is expected that the book will be mostly useful for text-mining experts, in particular for those not familiar with the health and life sciences, who will find several useful references to improve their knowledge of biomedical challenges related to text mining.

References

Abdou, S., & Savoy J. (2008). Searching in Medline: Query expansion and manual indexing evaluation. Information Processing Management, 44(2), 781–789.

Aronson, A.R., Demner-Fushman, D., Humphrey, S.M., Lin, J.J., Ruch, P., Ruiz, M.E., et al. (2005). Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. TREC.

Capurro, R., & Hjorland, B. (2003). The concept of information. In B. Cronin (Ed.), Annual Review of Information Science and Technology, Vol. 37 (pp. 343–411). Medford, NJ: Information Today.

Krallinger, M., Valencia, A., & Hirschman, L. (2008). Linking genes to literature: Text mining, information extraction, and retrieval applications for biology. Genome Biology, 9, Suppl 2, S8.

Mottaz, A., Yum, Y.L., Ruch, P., & Veuthey, A.L. (2008). Mapping proteins to disease terminologies: From UniProt to MeSH. BMC Bioinformatics, 9, Suppl 3,

Natarajan, J., Berrar, D., Dubitzky, W., Hack, C., Zhang, Y., DeSesa, C., et al. (2006). Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. BMC Bioinformatics, 7, 373.

Singhal A. (2001). Modern information retrieval: A brief overview. IEEE Data Engineering Bulletin, 24(4), 35–43.

Taleb, N.N. (2007). The black swan: The impact of the highly improbable. New York: Random House.

van Rijsbergen, C.J., & Lalmas, M. (1996). Information calculus for information retrieval. Journal of the American Society for Information Science and Technology, 47(5), 385–398.

Patrick Ruch, PhD
Bibliomics and Text Mining Group
Library and Information Sciences Department
Haute Ecole de Gestion
University of Applied Sciences Geneva
Carouge, Switzerland

Published online XXX in Wiley InterScience (www.interscience.wiley.com).
10.1002/asi.21236

Author Queries

AQ1: Citations were numbered; changed to conform to alphabetical journal style; please check carefully.
AQ2: Added definition of SNP; correct?
AQ3: Please provide volume and pages. If these are published proceedings, please provide the full name of the conference/workshop, page numbers, and publisher information.