Literature Analysis: A Version of Formalized Crowd Sourcing

John M. Weiner, Dr.PH
Senior Editor
Idea Analyses
Center for Study of Scientific Ideas
USA

Debora S. Bartoo, PhD
Adjunct Faculty
University of Denver
Strategic Innovation and Change
South Josephine St
Denver, Colorado

Abstract

This report explores the effectiveness of a formal version of crowdsourcing in assessing manuscripts submitted for publication. The credibility of the results may be questioned because the expertise of those responding may be unknown. Literature Analysis is an attempt to correct this potential defect in crowdsourcing by capturing the reviewed and refereed, published ideas from subject specialists. The feasibility of Literature Analysis in providing a mechanism for assessing the contribution of a proposed manuscript is illustrated using examples from two different subject areas – business innovation and dog-related disease. If there is merit to considering the world's author-specialists and their ideas, Literature Analysis offers a formalized process in the capture and use of these data.

Keywords: Ideas, crowdsourcing, data, repository, literature, analysis, peer, assessment

1.Introduction

Numerical databases have demonstrated success in the study and understanding of quantitative relationships. The rigor by which the databases are constructed and utilized underscore the credibility of findings based on the retrieved data.

Textual databases are different. The structure and use of a digital repository reflect an earlier technology based on the management of paper. Documents are stored for retrieval as units. Once retrieved, the digital document usually is explored using techniques similar to those used with paper. This process is advantageous by being familiar, unchanged from earlier times. It also is disadvantageous by impeding efficient and effective processing of the text.

1.1. CrowdSourcing

Crowd sourcing is a recent attempt to acquire information or accomplish work by recruiting a large number of participants, often from an online community. (Lloret 2013) By tapping into the common knowledge pool, a consensus version of the topic can be rapidly acquired or new solutions considered. However, the accuracy or completeness of the acquired information may be difficult to document as the crowd is anonymous with respect to qualifications for supplying the information.(Cummings 2013)Literature Analysis (Weiner 1979, 2011) capitalizes on the value of obtaining information from a large number of individuals. The difference is that each respondent is a documented specialist in the topic. Participation by these experts is complete by capturing their peer-reviewed and refereed, published ideas. Using those data, consensus descriptions can be constructed. In addition, by focusing on newly emerging ideas, research strategies can be developed.

1.2. Dealing with the Actual Volume of Ideas

The Literature Analysis process is not without potential problems. There is a significant drawback associated with identifying, extracting, and organizing the authors' ideas contained in documents.

That is the larger volume that must be considered. Instead of applying the concept of Occam's Razor dealing with the **smallest, most effective set of ideas,** the Literature Analysis approach is designed to deal with the **millions of ideas actually presented by the authors**. Some will be irrelevant and represent those fading from interest. Others will be infrequent as they are emerging from the single author's presentation to a larger number of individuals describing the idea. Still others will be observed frequently, at any time, representing a type of consensus.

There also are significant advantages associated with the Literature Analysis approach. One is the ability to employ quality control procedures in the construction of the idea database as well as in the accomplishment of the higher cognitive functions.(Bloom 1956, Chen 1988, Hoffman 1980, Weiner 1979) That is, the intellectual tasks required can be made more transparent, thus enhancing the actual procedures used by the specialist. This oversight can be beneficial to the student who is challenged to accomplish the transformation to professional. Each function can be subdivided into three key areas, namely, the development of – measures/observations describing the phenomenon considered, criteria using those variables, and decision-rules leading to resulting actions.

1.3. Reconsidering the Tower of Babel

The biblical story of the Tower of Babel implied that different languages prohibited effective communication. As a result, specific disciplines (e.g., mathematics, statistics) became substitutes to minimize the effects of the polyglot. Text mining (Hearst 2003), involves application of specific algorithms in the analysis of text. That methodology forced reconsideration of differences and similarities between languages.Irrespective of the method of delivery, each language is made up of terms combined to present thoughts (i.e., ideas) and does so by presenting those ideas in sentences.

This consistency is illustrated by considering examples of analyses from two different disciplines – business innovation and dog-related diseases. The individuals working in those areas adopt different writing styles and emphasize different ideas, concepts, and issues. The question of interest is – ***Can formal identification and organization of the ideas presented by specialists in their scientific documents be used to effectively assess the contributions of a new document?*** This version of crowd sourcingcould be useful in the preliminary screening of proposals. Using the ideas as an evaluative basis, four groups of documents could be recognized. These are:

1. Documents presenting new ideas that enhance understanding of the phenomena.
2. Documents presenting confirmatory evidence of existing ideas and concepts in a new environment.
3. Documents presenting new methods providing improved insights.
4. Documents presenting existing ideas in previously studied environments.

Of the four categories, the first three provide relevant information and could be disseminated without further review. The fourth category would require expert assessment in order to determine the contribution made.

2. Methods

The Literature Analysis method is based on the premise that software can accomplish the following tasks.

1. Identify and separate the authors' sentences so that each can be considered as a domain containing informative terms (nouns, adjectives, or gerunds) combined by the authors to present thoughts or ideas.
2. Identify each pair of informative terms (i.e., idea) within each sentence.
3. Prepare a data record consisting of the idea and the bibliographic data describing the document and its location.
4. Organize those data records as a file for subsequent use.

The software identified over 3.6 million ideas in the dog-disease documents for the period 1980 – 2013. The business innovation documents for 2013-2014 provided 24,151 ideas. The software identified informative terms using defining suffixes and contextual meaning of terms. The median number of ideas contained within a sentence was 8 (1 to 10). The time required to identify, extract, and organize the ideas was 0.3 minutes and involved three readings of each document. The first identified informative terms. The second prepared the idea records. The third confirmed these records.

The software identified a median of 85% of the informative terms (66% to 99%) employed by the author-specialists. Over 95% of the ideas were correctly identified and included in the idea database. The identification of informative terms compared favorably with previous approaches – human indexing (50% - 70%), statistical text mining methods (30% - 50%), and random selection (10% - 20%). The processing time was significantly less than the previous methods including text mining designed to cluster documents into themes. That approach required manual inspection of the clusters to validate document membership.

Exhibit 1 shows an example of the identification, extraction, and organization process. The particular document was randomly retrieved from the Pub Med Bibliographic Database. The topic considered was dog-related diseases and the document was entered into Pub Med during 2014. The idea records show the pairs of these informative terms together with the location of the sentence containing the ideas. The identification number was assigned by PubMed and provides a portal for rapid retrieval of the document of interest.

Exhibit 1: Example of Sentence Containing Ideas and the Ensuing Data Records Included in the Idea Repository.

Source

Mahmoudvand H[1], FasihiHarandi M[2], Shakibaie M[3], Aflatoonian MR[4], ZiaAli N[5], Makki MS[6], Jahanbakhsh S[1]. Scolicidal effects of biogenic selenium nanoparticles against protoscolices of hydatidcysts. Int J Surg. 2014;12(5):399-403. doi: 10.1016/j.ijsu.2014.03.017. Epub 2014 Mar 28.PMID: 24686032

Abstract

Cystic echinococcosis (hydatid cyst, CE) as a zoonotic parasitic infection caused by the larval stage of the dog tapewormEchinococcusgranulosus is still an important economic and public health concern in the world. One of the treatment options for CE is surgical removal of the cysts combined with chemotherapy using albendazole and/or mebendazole before and after surgery. Currently, many scolicidal agents, which have some complications, have been used for inactivation of the cyst contents. Therefore the development of new scolicidal agents with low side effects and more efficacies is an urgent need for surgeons. The present study was aimed to investigate the in vitro scolicidal effect of selenium nanoparticles biosynthesized by a newly isolated marine bacterial strain Bacillus sp. MSh-1 against protoscoleces of E. granulosus. Protoscolices were aseptically aspirated from sheep livers having hydatid cysts. *Etc.*

Idea Records

Primary	Related	Year	Ident	Sentence
albendazole	chemotherapy	2014	24686032	2
albendazole	mebendazole	2014	24686032	2
aseptic	aspirate	2014	24686032	8
aseptic	hydatid	2014	24686032	8
aseptic	liver	2014	24686032	8
aspirate	aseptic	2014	24686032	8
aspirate	hydatid	2014	24686032	8
aspirate	liver	2014	24686032	8
bacillus	bacterial	2014	24686032	5

3. Results

3.1 Dog-Related Ideas

A randomly selected document from the 2014 set entered in PubMed provided ideas (see Exhibit 1). Those ideas were tracked throughout the period 1980-2013.

Table 1: Ideas from Document 24686032. Frequency of Occurrence through Time.

Primary	Related	Total	1980-84	1985-89	1990-94	1995-99	2000-04	2005-09	2010-13
Dog	Infect	4545	412	491	775	751	897	1167	52
Dog	Health	2240	76	158	244	354	402	861	145
Dog	echinococcus	229	13	31	39	22	25	79	20
Health	Infect	199	7	9	22	35	29	81	16
echinococcus	Infect	125	11		23	18	25	42	6
Dog	Parasitic	117	2	6	6	9	10	45	39

Table 1 shows an excerpt from that analysis. The most frequently occurring ideas are shown for each five year period. The core ideas (i.e., highest frequency through time) were respectively, dog with infection and health. The moderate ideas showed frequencies 43 to 229. Ideas in this grouping included: **echinococcosis** with respectively, health, hydatid, and tapeworm. The low frequency ideas ranged from 1 to 30 and included the combination of albendazole and mebendazole (frequency = 4). The new ideas supplied by the authors were not used in previous reports and were recorded as 0 frequency. Those are given in Table 2.

Table 2: New Ideas Considered in the 2014 Document (PMID 24686032

Primary	Related
albendazole	chemotherapy
aseptic	Aspirate
aseptic	Hydatid
aseptic	Liver
bacillus	Selenium
bacterial	Selenium
chemotherapy	mebendazole

Figure 1: Complete Idea Structure Considered by the Authors of Document 24686032

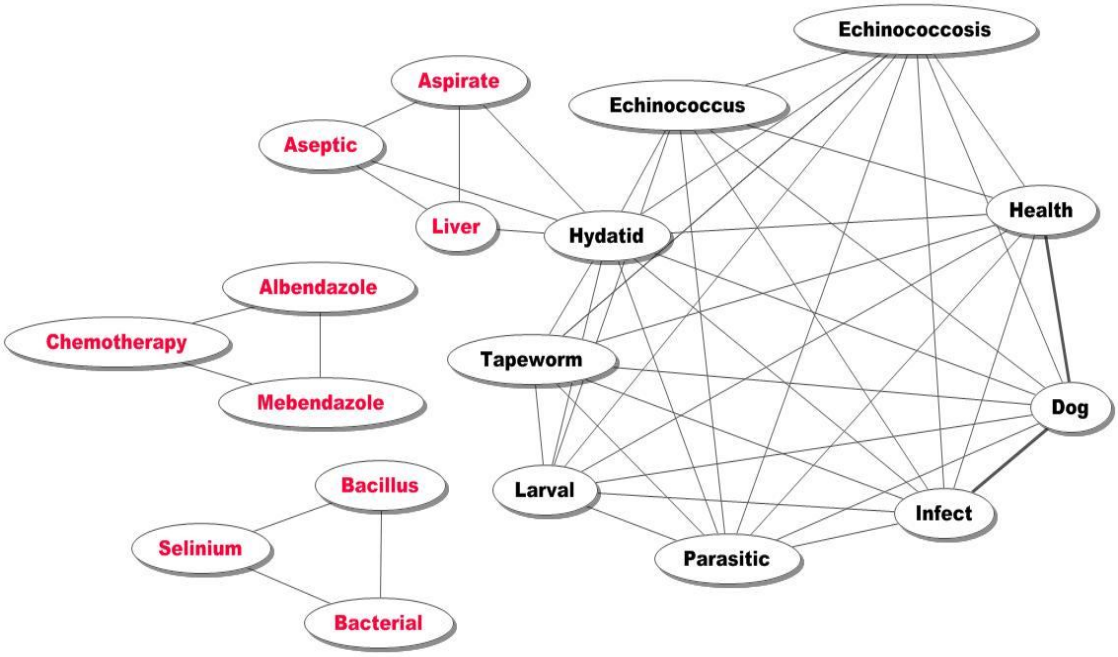

Figure 1 shows the combined idea structure consisting of ideas considered by specialists throughout the period 1980-2013 plus the new ones presented by the authors of the 2014 article. The consensus ideas are shown in black and the 2014 ideas in red.

The informative terms are shown as nodes and the ideas representing pairs of these terms are shown by the links between the nodes. The structure suggests that the authors went beyond previous studies by considering therapy. To do that, they linked two forms of chemotherapy with a bacterium obtained from the liver. The link between the consensus and new ideas is via the hydatid node. As expected when the idea structure is expanded, new questions emerge. The authors recognized that and suggested in their concluding sentence that – "However, the in vivo efficacy of these NPs remains to be explored."

3.2 Business Innovation

In assessing the efficacy of Literature Analysis in dealing with information of this type, documents from the business innovation literature for 2013 and 2014 were analyzed. They differed from those from other disciplines by including blog news reports in addition to more traditional scholarly documents. In addition, only two years were considered. The relatively brief period reflects the focus on new issues and products.

Table 3: Terms in Random Article Compared with Use as Ideas in 2013 and 2014.

Term	Random	Sum	2014	2013	Term	Random	Sum	2014	2013
customer	8	1236	273	963	authentic	8	72	5	67
solution	2	456	0	456	project	4	52	0	52
Google	2	438	81	357	leader	3	43	0	43
create	8	267	61	206	collaboration	11	35	0	35
ability	8	233	26	207	organization	3	35	0	35
change	3	217	56	161	inspiration	8	29	29	0
process	3	187	0	187	learning	4	23	0	23
innovation	22	162	46	116	davidson	3	14	6	8
important	9	156	11	145	possibilities	8	12	12	0
executive	9	154	54	100	conversation	8	11	5	6
growth	3	135	40	95	finding	6	10	3	7
insight	4	116	45	71	question	5	7	0	7
people	10	108	0	108	navigate	4	3	0	3
challenge	10	101	20	81	culture	6	2	0	2

Table 3 shows the terms from the randomly selected article and the frequency of those terms in ideas cited in 2013 and 2014. The terms are arrayed using the sum for the two years. The number of times each term was included in an idea in the random article also is given. The total frequency of each idea is compared with its use in the random article. Terms may be classified as general in meaning (e.g., customer) or specific (e.g., Google). General meaning terms, as expected, occurred with high frequency. The top two were customer and solution. The most frequently used ideas in the random article involved the term – innovative. That term was used in 22 ideas. The frequency of use in the 2013-2014 document set was relatively modest (162 times with the majority in 2013). In contrast, the other specific term of interest – inspiration – was used 8 times in the random article and only 29 times in 2014. It was not used in ideas in 2013.

Exhibit 2: Sentence and Idea Records from Business Innovation Document 207501.

Senior executives consider innovation one of the top threedrivers of growth in the next 3 to 5 years (Barsh, Capozzi, &Davidson, 2008).

Primary	Related	Year	Ident	Sentence
davidson	executive	2014	207501	1
davidson	growth	2014	207501	1
davidson	innovation	2014	207501	1
executive	davidson	2014	207501	1
executive	growth	2014	207501	1
executive	innovation	2014	207501	1
growth	davidson	2014	207501	1
growth	executive	2014	207501	1
growth	innovation	2014	207501	1
innovation	davidson	2014	207501	1
innovation	executive	2014	207501	1
innovation	growth	2014	207501	1

The focus on general meaning terms may reflect the lack of a discipline-specific set of terms associated with business innovation. In that case, general terms – change, ability, create, growth, etc. – may be considered as relevantly descriptive. The inclusion of many of these terms in frequently occurring ideas suggests this possibility.

4. Discussion

Given the exponentially increasing amount of information available, the individual, student or professional, is faced with the challenge of identifying, extracting, organizing, and utilizing facts and opinions to form new descriptions of a topic. Research plays an important role in dealing with the deluge of information. In a real sense, research is a formalized learning process, pushing back the unknown and discovering aspects of new topics. Many books have been written describing research methods. Few however have focused on the full range of tasks and how they can be performed using the most effective capabilities. Those tasks are:

1. **Developing a learning resource**. This repository of essential data provides the ingredients needed to construct descriptions of the topics and to develop new strategies leading to advancing knowledge. This construction process can best be accomplished by software designed to identify, extract, and organize specific data from the scientific publications describing the topic. By using software, there is enhanced transparency of procedure. The measures, criteria, and decision-rules must be specified in advance and use of those performed in a consistent fashion.

2. Using the learning repository. The data in the learning resource represent the totality of information regarding the topic. The student can rapidly learn the basic description of the topic by focusing on the information considered most relevant by the subject specialists who developed it. This specialist-guided mentoring and instruction expands on the work of an author of a textbook by providing a worldwide view of the topic rather than an individual's interpretation.

The professional can rapidly access the subsets of data considered relevant in addressing a specific topic. In addition to saving time and effort, the subsets selected will be more complete than an individual's personal perspective of the knowledge, thus providing more opportunities to expand the body of relevant information.

3.Formalized learning. With the learning repository available, the individual's time and energy can be shifted from acquiring data to using the higher cognitive functions. The first of these is synthesis, the construction of new descriptions of the topic or issues in the topic based on arrangements of the essential data from the learning repository. In contrast to previous methods, this version of synthesis would lead to many different possible arrangements. This array of new descriptions offers the opportunity to compare, evaluate, and judge each leading to a ranking denoting most applicable to least.

The selected application could be called a testable hypothesis or the question to be studied in a research program. It is the ***best*** arrangement of the existing facts and represents the most plausible portal to new understandings. Those results will be determined by the new study.

In this sense, the process of developing a study hypothesis and the process involved in populating that hypothesis with new information are transparent, quality-controlled events. Learning accomplished using this approach can be described as a system of formalized performance of specific tasks, operationalized critical thinking, or simply as research.

4.1 Literature Analysis and Crowd Sourcing

Literature Analysis is an attempt to capture the observations, perceptions, and opinions of the subject specialists as a group rather than the current practice of selecting two or three as representatives of the total. To do this, different processing methods must be employed. A major difference is the focus on managing the full set of ideas presented by the subject specialists. By separating the management methods to best accomplish the objective of the task, a significant shift in focus and energy is possible. The traditional cognitive functions could be separated into those best performed by software, technicians, or subject specialists. The latter are best utilized by directing their attention to the functions dealing with comparison of syntheses and the ultimate choice of a particular synthesis for study.

Making those intellectual functions more transparent is an important step in enhancing critical and creative thinking.

By formalizing the utilization of information from the entire group of subject specialists, via their published ideas, the advantages of crowd sourcing are possible while ensuring that the information is provided by credible authorities. In addition to the advantages of establishing a consensus on demand, temporal analyses are feasible showing the dynamics of emphasis on specific ideas. This ability is important in enhancing identification of interesting new strategies for expansion of the knowledge dealing with a specific topic. While these results are possible using paper-oriented methods, the time and effort are considerably longer and the process more private. Those conditions make learning of effective information utilization slower and more difficult.

5. Summary

With the tools described above, the expert is needed to clarify and expand the higher cognitive functions while the more clerical/mechanical ones are converted to a transparent, evidence-based system. By making the analysis transparent and quality-controlled, the shift to true intellectual prowess by the expert is facilitated. In a similar fashion, the student, new to the subject, can begin learning by solving problems and by building new idea structures. The ability to acquire, organize, and utilize the ideas enhances the transformation from novice to professional. The need to spend long hours in the library stacks is replaced by a need to spend time thinking and researching. Those actions could yield an operational description of critical and creative thinking.

With the ideas organized and usable, students at all levels can build numerous idea structures representing a given topic. This array shifts the focus from a single answer to a spectrum of possibilities each with desirable and undesirable characteristics. Weighing those attributes and developing rationales is an example of the critical thinking process. Translating those functions to transparent, quality-control procedures is an example of the Literature Analysis approach. If there is an advantage to capturing the observations, opinions, and perceptions from subject specialists and to organize those for rapid utilization, the methods must change from those based on management of whole units of text to those focused on the actual content of those units. Literature Analysis involves the identification, extraction, organization, and utilization of important elements making up text, namely, the authors' ideas. The captured data offer the ability to benefit from the subject specialists in a formalized manner.

References

Bloom, B. S.; Engelhart, M. D.; Furst, E. J.; Hill, W. H.; Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals*. Handbook I: Cognitive domain. New York: David McKay Company.

Chen J. The Natural Structure of Scientific Knowledge: An Attempt to Map a Knowledge Structure. Journal of Information Science 1988;14:131-139.

Cummings, S., Daellenbach, U., Davenport, S., & Campbell, C. (2013). "Problem-sourcing": A re-framing of open innovation for R&D organisations. Management Research Review, 36(10), 955-974. doi:http://dx.doi.org/10.1108/MRR-07-2012-0177

Hearst M. What is Text Mining? <http://www.sims.berkeley.edu/~hearst/text-mining,html, 2003.

Hoffman E. Defining Information: An Analysis of the Information Content of Documents. Information Processing and Management 1980;16:291-304.

Lloret E, Plaza L, Aker A. Analyzingthecapabilitiesofcrowdsourcingservicesfortextsummarization. LangResources&Evaluation(2013)47:337–369DOI10.1007/s10579-012-9198-8

Weiner JM. Issues in the Design and Evaluation of Medical Trials.G.K. Hall, Boston, MA 1979.

Weiner JM. Effective Creativity in the Workplace Lambert Academic Publishers, Germany, (2011)http://amazon.com.