Medicinal chemistry

Uguroglu et al. Medchem 2011, 1:1
DOI: 10.4172/2161-0444.1000e101

Editorial Open Access

Identification of Potentially Relevant Citeable Articles using Association Rule Mining

Selen Uguroglu[1], Oznur Tastan[2], Judith Klein-Seetharaman[1,2]* and Sanford H. Leuba[3]*

[1]*Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213*
[2]*Microsoft Research New England, Cambridge, MA 02142*
[3]*Departments of Cell Biology and Bioengineering, University of Pittsburgh Schools of Medicine and Engineering, Hillman Cancer Center, UPCI, Pittsburgh, PA 15213*

Abstract

Due to the increasingly larger and more interdisciplinary nature of scientific reporting, it is becoming more difficult to identify all the potentially relevant, citeable articles in reference lists of publications such as scientific papers, reports, grant proposals and patent applications. Authors may miss and/or give inaccurate citations, potentially hindering progress in a discipline and on a personal level, and change the importance and impact of an investigator's work. Given the emphasis on quantitative means for assessing productivity, including the number of literature citations, efforts are needed to assist authors in the identification of potentially relevant articles to cite. Prior work has analyzed citation network structure and characteristic features and correlated these with other variables, such as country of origin, journal impact factor and open access status. As a result, problems have been revealed, such as underrepresentation of third-world countries, a high incidence of self-citation, and unsystematic quotation habits in review articles. With the exception of gross plagiarism detection software, however, no attempt has been made to develop a practical solution to identifying potentially relevant, citeable articles that may have been missed. Here, we use statistical methods to help in the retrieval of relevant literature from existing publications. Specifically, we exploit the fact that publications reporting specific findings are typically quoted together as grouped-co-citations in their respective contexts. Our approach can automatically construct rules for co-citation by automatically extracting co-citation overrepresentations in manuscripts. This approach should help authors and reviewers identify potentially relevant, citeable articles.

Introduction

Scientists rely on a trust-based system – the peer-review system – for publications and grant proposals. When evaluating scientific publications and grant proposals, reviewers assess the accuracy of citations of prior work to place the present work into context. Often, a reviewer points out a missed reference that should be included, or a paper is rejected based on its inaccurate or incomplete overview of the relevant literature. The peer-review system works quite well because scientists will generally do the best job they can since a scientist's reputation and the acceptance of the paper depend on it. However, the increasingly interdisciplinary nature of biomedical research involving collaborators from diverse disciplines and the shear volume of journals and published communications make it increasingly difficult to fully assess the accuracy of citations, which leaves opportunities for scientists to miss citing relevant publications. Currently, the only tools available to aid scientists in creating accurate reference lists are the search engines used to mine literature databases such as Medline, Web of Science, Google Scholar and Scopus or more general search engines such as Google. Manual inspection of reference lists of relevant papers, targeted publication retrieval based on citations in relevant original papers or reviews is an integral part of assemblage of the creation of a reference list and placement of the own work into context of prior work. However, after a reference list has been created, there are no methods to systematically assess the completeness or accuracy of the reference list. Additional assistance in identifying potentially relevant, citeable articles that an author might not be aware of would be a major asset to anyone preparing a scientific document. Such tools would be particularly valuable to support collaborative efforts, in which scientists from disparate fields come together to work on new application areas at the boundaries between their respective areas of expertise.

On the other hand, there could also be intentional cases of missed citations because there are many non-scientific reasons scientists may have to omit or not properly cite prior publications. Such behavior may come from a desire to enhance the weight of contribution by a senior author and downplay the significance of other reported data. Unscrupulous scientists may even omit key literature that is at odds with their hypothesis. Thus, an urgent need exists to develop quantitative and less-biased tools to help scientists in their efforts to create and evaluate the accuracy of reference lists.

Here, we describe a new approach to assess completeness of a given reference citation list based on statistical analysis of co-citations of references cited in other publications. First, we manually assembled a small dataset of publications which included a subset of papers that does not cite a particular paper although that paper should have been cited in the given context of the topic reported on. We then extracted all the references and the occurrences of the references in the text from the publication files in portable document format (PDF) automatically and statistically evaluated co-occurrence of citations in these papers. The results validated the manual approach by suggesting the relevant publication omitted in some of the reference lists as a citeable article. We then tested the approach on an unknown, larger dataset of publications in the Proceedings of National Academy of Sciences of the United States of America (PNAS), and manually inspected the top-ranked list of suggested co-citations. In all cases, the suggested co-citations were highly related to each other and their co-citation was justified. This

***Corresponding authors:** Sanford H. Leuba, 5117 Centre Avenue, 2.26a Hillman Cancer Center, Pittsburgh, PA 15213, Tel: 412-623-7788; Fax 412-623-4840; Email: leuba@pitt.edu

Judith Klein-Seetharaman, Biomedical Science Tower 3, Rm. 2051, 3501 Fifth Avenue, Tel: 412 383 7325; Fax: 412 648 8998; Email: jks33@pitt.edu

Received December 01, 2011; **Accepted** December 01, 2011; **Published** December 03, 2011

Citation: Uguroglu S, Tastan O, Klein-Seetharaman J, Leuba SH (2011) Identification of Potentially Relevant Citeable Articles using Association Rule Mining. Medchem 1:e101. doi:10.4172/2161-0444.1000e101

result demonstrates that the idea of retrieving co-citations is useful in identifying articles relevant to a particular topic, and pointing out when such relevant articles are missing from a given reference list.

Methods

Data

In this work we evaluated our approach on two datasets: a small dataset that is collected by an expert, and a large dataset that is obtained by downloading publications in PNAS within the years 2000 and 2001. The small dataset contains 12 papers taken from a specific domain that is closely related to the expert's field of interest. The large dataset consists of around ~3170 papers on a variety of topics. Papers are retrieved automatically with a python script from PNAS archives (http://www.pnas.org/content/by/year). Some of the papers had to be disregarded due to limitations in extracting and parsing the references section from the papers. How we extracted references is explained in the next section.

Reference extraction

Reference fields were extracted from PDF files as text using python scripts based on heuristics such as end of file, author/data/journal format for the actual references and numbers in brackets etc. for the citations in the text. Each reference was identified by Medline reference PMIDs and stored in a database including metadata (authors, date, journal, title). All references were extracted in association with a sentence or paragraph.

Algorithm

We applied the association rule mining method to discover interesting relationships between publications' co-citations. A rule in this context defines which publications should be co-cited. $p_1 ==> p_2$, if p_1 is cited, p_2 is likely to be cited together with p_1. Association rules were discovered by using the Predictive Apriori algorithm [1] implemented in the Weka toolkit. This algorithm finds the *n* best association rules which maximize the resulting criterion by dynamically pruning redundant rules that cannot contain better solutions than the best ones found so far. We identified rules that describe pairs of articles co-cited, but in the analysis we analyzed in particular those pairwise rules that linked three papers together through the respective pairwise rules.

Results

Manual inspection of a small method development dataset

First, we developed a small set of publications for which we manually derived rules of co-citations of publication. In Table 1, the incidence of co-citation of 3 seminal papers (all from journals with impact factors >10) in the same field is shown, together with the group affiliation. Co-citations are those that occur together in the same sentence. In all 11 out of 12 papers, the three papers are cited together. The only exception is laboratory Group #1's 2006 paper, which does not cite Group #2's 2001 paper. Given that the vast majority cite all three papers suggests that this omission may potentially be a mistake. Thus, based on manual inspection of these publications, it appears that Group #2's 2001 paper is a missing reference that should have been included in Group #1's 2006 paper. The evidence is co-citation of this reference in other publications related to the topic.

Automatic retrieval of the rules for the small development test set

Next, we wanted to test if co-citations can be exploited as a feature to discover potentially missed citations automatically. Our goal was to identify cases of missed citations such as the one described above, where many papers co-cite two or more references, so that if in a given reference list, these papers are not co-cited, the authors of that list could be alerted to the fact. This would then provide a practical means for these authors to check if the predicted article should be included or not. To test our hypothesis of co-citations being a useful feature to suggest potentially citeable articles, we used 12 publications of the above small set of 15 publications to find if we could discover the manually identified three-way co-citation rule automatically. To this end, we conducted association rule mining using the Weka toolkit (see Methods). First, we extracted references from publication PDF files using heuristics such as sequential number of new paragraphs and the appearance of reference indicating keywords. The extracted references were matched to PubMed identifiers to ensure unambiguous retrieval of the cited reference in each case. The resulting meta-data (author, date, title, journal name, page numbers) were stored with their PubMed identifiers in a database. Co-occurrences of references at the end of single sentences were counted in all 12 PDF files. A threshold value of 0.7 retrieved the co-occurrence of the three references. The 3-reference rule was identical to the one described above that had been identified by human knowledge. Above this threshold was one additional rule, for the co-occurrence of two other references, papers 15695630 and 11114182 (Table 2). The novel 2-reference rule was then inspected by the human expert. Feedback from the human expert confirmed the validity of this automatically discovered rule, which suggested co-citation of these two highly related papers.

Automatic retrieval of rules for a large database of publications

Having demonstrated that we can automatically retrieve known

	Group #1 2000 paper (10618382)	Group #2 2001 paper (11427891)	Group #3 2002 paper (11854495)	(Medline PMID)
Group #2 2003 paper	Yes	Yes	Yes	12522259
Group #5 2003 paper	Yes	Yes	Yes	12831877
Group #6 2003 paper	Yes	Yes	Yes	12897855
Group #7 2004 paper	Yes	Yes	Yes	15321707
Group #8 2004 paper	Yes	Yes	Yes	15447507
Group #3 2005 paper	Yes	Yes	Yes	15663933
Group #9 2005 paper	Yes	Yes	Yes	16002089
Group #7 2005 paper	Yes	Yes	Yes	15882698
Group #10 2006 paper	Yes	Yes	Yes	16453064
Group #1 2006 paper	Yes	***No***	Yes	17043216
Group #11 2006 paper	Yes	Yes	Yes	17012315
Group #6 2007 paper	Yes	Yes	Yes	17108322

Table 1: Co-citations of three publications in a small list of related publications. Incidence of grouped-co-citation of 3 seminal papers in a particular field by 11 independent laboratories. Ten out of 11 groups cite the 3 seminal publications in a grouped-co-citation.

Grouped co-citations (PMID)			Accuracies
10618382	11854495	11427891	0.26415, 0.25931, 0.21341
11854495	11427891	10618382	0.25931, 0.23263, 0.18464
11854495	11427891	15695630	0.25931, 0.12725, 0.17077
11114182	15695630	11427891	0.23597, 0.21610, 0.21341
11427891	10618382	11114182	0.23263, 0.13293, 0.07933

Table 2: Automatic retrieval of grouped co-citations obtained manually in (Table 1). The first three columns in each row give the PMID of each publication involved in a rule and the final column gives the accuracies for the association rules in that row. For each set of three papers, there are three accuracies reported (last column), the first one is for rule paper 1 ==> paper 2, second accuracy is for paper 2 ==> paper 3 and third one is for paper 1 ==> paper 3. Paper 1,2,3 refers to the columns. Id's are PubMed Ids. The above three papers identified manually are highlighted in green: Group #1 2000 paper (10618382), Group #2 2001 paper (11427891), Group #3 2002 paper (11854495).

Grouped co-citations (PMID)			Accuracies
9153396	9153395	9450543	0.89993, 0.24983, 0.74998
10676951	10521349	11207349	0.89993, 0.74998, 0.23809
10676951	10521349	10963602	0.89993, 0.25926, 0.38095
10676951	10521349	10952317	0.89993, 0.22222, 0.33333
10676951	10521349	11385503	0.89993, 0.11004, 0.14223
9521922	9521923	9521921	0.85710, 0.50000, 0.70000
10089887	10069338	10458908	0.85710, 0.40000, 0.74998
8207839	8057491	1387031975	0.85710, 0.33325, 0.74998
8479522	8479523	9039259	0.83330, 0.24983, 0.37494
8479522	8479523	8247009	0.83330, 0.24983, 0.37494

Table 3: Top 10 best rules for the PNAS dataset.

rules, and identify new ones we did not consider previously, from the small set of manually curated dataset, next, we applied the same algorithm to a larger database of publications. We downloaded two years of publications from the Proceedings of the National Academy of Sciences USA (PNAS) and applied the Predictive Apriori algorithm to discover rules of co-citations. The top 10 most highly ranked related papers are listed in (Table 3). We manually inspected all of the retrieved publications linked by co-citation rules, and found that retrieval was justified in all cases. For example, the first rule (9153396, 9153395, 9450543) is a set of papers on regulation of p53 stability by Mdm2. The first two papers are published back to back in a single issue of Nature, while the third was published independently a few months later in FEBS letters. The next rules (lines 2-5 in (Table 3)) link two papers 10676951 and 10521349 to other publications. 10521349 is a famous paper describing the use of a transcriptomic dataset on distinguishing two different types of cancers, namely AML and ALL, by classification. Similarly, 10676951 describes classification of different B-cell malignancies through microarray data analysis. These two publications inspired hundreds of studies on improving cancer classification algorithms. The different subsequent rules all involve more recent papers using the same approach on different datasets. Thus, a researcher would need to consider citing 10676951 and 10521349 together due to their pioneering role in this field, while inspection of the other rules may lead to identification of papers based on the relevance to the particular type of cancer under study for example.

Discussion

The manner in which scientists cross-reference each other's work results in a complex network of citations. Such citation networks have been studied previously [2,3,4], especially with the purpose of using them as reporters of collaboration networks [5]. Citation networks have been useful for investigating how collaborations develop and are sustained. Citation networks have also been investigated to study the evolution of scientific discoveries. For example, it was proposed in 1964 to use citation networks to report on the history of science [2]. Garfield has since shown that citation networks are invaluable for the study of the history of science and has developed software to produce "historiographs", visual representations of citation history, that provide scientists with an opportunity to browse the network of citations starting with a specific seed publication or a keyword or author [6,7]. A field closely related to citation network analysis is also that of scientist networks, such as editors [8,9] and authors [10] and references therein). While citation and scientist networks are distinct (in one the nodes are papers, in the other the nodes are persons), they are both networks with related features, and most recently citation networks and author networks have been shown to co-evolve [4]. While much work focuses on correlations of the author citations or editorships with country of origin [8], institution, or degree of collaboration and its social implications (*e.g.*, papers published by less affluent countries [11-14]), author co-citation analysis can be practically useful. For example, if author X and author Y are often cited together, this could be useful information for scientists new to the field in which X and Y are working [15,16]. Ideas to exploit the revelation of ethical problems in citation practice are emerging, for example, it was shown that blinded peer review improves review quality [17] and, Andrews proposes that author co-citation analysis (author's names appearing in the same reference list) "helps identify the most productive and prominent authors in the field, the amount they are cited, the amount they are co-cited […] and the authors who appear to work in similar subject areas." However, there currently is no effort to help scientists identify co-cited *articles* as opposed to scientists.

Citation and author network analysis has been useful in detecting citation bias. The networks naturally partition into topics and reflect the bias of authors to cite recent papers and papers they have read; these properties have been captured in the so-called TARL (topics, aging and recursive linking) model [4]. TARL takes all three types of bias in citation habits into account. The citation preference for read articles is, for example, also reflected in the "open access advantage" [18,19]. A large literature on highly cited papers derives from practical listings of "what articles should a scientist in field X have read" [20-25] and from reports showing the bias in referring to highly cited papers ("the rich-get-richer phenomenon" [26,27,28]. Copying behavior may also contribute to the typical network properties [29], and authors like to cite themselves [30]. The citation bias manifests itself in higher frequency of citations from high-impact journals [31,32,33] and university affiliation [34], although cause and effect are less clear in these correlations. These behavioral patterns – citing papers that have already been cited frequently, citing your own papers, citing papers of friends, having friends cite your papers, citing papers that are easily accessible, and citing only recent papers (*i.e.,* not paying attention to older work and thus reinventing the wheel) - often occur inadvertently without clear intent of unethical behavior. The availability of large datasets of citations has allowed global analysis of the features of citation networks. For example, it was shown that citation networks like many other biological and social networks are characterized by features such as the "small world effect", in which the average distance between nodes in a network is small; the degree distribution, which is often skewed; grouped-co-citations, for example, if person X has two friends Y and Z, it is likely that Y and Z are also friends; and community structure [35].

Here, we propose that we can exploit the network structure of citations to search for patterns in citations, and use deviation from established patterns to reveal potential outliers. The pattern we propose to use is co-citation, referring to the scenario where two or more papers are cited together at the end of statements in publication texts. We

propose that outliers in co-citation patterns can be used to identify intentional or unintentional omission of publications in reference lists and help reviewers and authors in critical assessment of a given reference list. We first developed a small set of related publications in which we had manually identified omissions of relevant articles, followed by automatic retrieval of these manually identified rules. Having successfully demonstrated the ability to retrieve these rules (and additional ones), we then expanded the approach to a large dataset of publications from PNAS, where we identified rules of co-citation that led to retrieval of clearly related publications. While a careful analysis of the literature would have probably resulted in identifying these sets of papers manually, our tool allows a complementary way that can lead to a speed-up in the discovery of related articles. Conceptually, our tool is thus useful not only in assistance in critical evaluation of reference lists, but also in assisting the reference list building. Search engines such as PubMed (www.ncbi.nlm.nih.ogv/pubmed) are beginning to provide such tools, going beyond the typical search for keywords, author names, cited references etc. For example, the "related citations" tool in PubMed, allows users to quickly skim through potentially related articles, based on word occurrences in the respective abstracts, titles and Mesh terms (http://www.ncbi.nlm.nih.gov/books/NBK3827/). This provides highly complementary information and is entirely different from the approach described in this paper. Here, we propose to use a new feature, namely co-citation, as a source of information with regard to relatedness. The complementary nature can be appreciated qualitatively in that in the analysis of the paper co-citation rules described in (Table 3) from the PNAS dataset, we did not find the highly co-cited papers cross-referenced through the top ranked "related citations" list in PubMed.

Conclusions

In this paper, we have described research to develop tools for use by the entire scientific community. We identified co-citation as a new feature to retrieve related publications and propose that this feature can be exploited in a practical way to reveal potential sources of citation bias quantitatively. Other features can be derived based on more complex language modeling and/or network analysis. For example, it has recently been proposed to use the entire list of references of a group of papers to find related publications [36]. This kind of approach may be implemented in our setting to improve finding potentially missed citations. Such approaches could be provided as novel capabilities in publication search engines, for example in NCBI, to incorporate this tool for widespread reduction of citation bias in the literature. We expect our approach to be useful for authors, reviewers (both manuscript peer review and grant proposal review), editors and readers. We expect that widespread use of the software generated by our proposal will help authors in literature citation.

Finally, we should note the limitations of our automated tool in terms of finding potentially relevant, citeable work in disciplinary or interdisciplinary areas. In studies of the distribution of papers judged relevant to a topic area -- the overlap between papers that are tagged as "on the topic" and papers that have the "right citations" is generally no more than 30% and can be much less. In other words, there are potentially relevant works that are recognized by indexers and not by citing authors and vice versa [37]. Don Swanson has written extensively about "logically but not bibliographically connected literatures" -- two bodies of work that, if brought together have a clear relationship to each other but that lack both indexing and citation links that would allow discovery. His major examples are Raynaud's disease and fish oil and migraines and magnesium [38,39,40]. The tool we are describing is unlikely to address this problem since it relies on citations.

Acknowledgements

We would like to thank Dr. Katherine W. McCain for discussion and Michelle L. Kienholz for critical reading of the manuscript.

References

1. Scheffer T (2001) Finding association rules that trade support optimally against confidence. Principles of Data Mining and Knowledge Discovery 2168: 424-435.
2. Garfield E, Sher IH, Torpie RJ (1964) The use of citation data in writing the history of science. Institute for Scientific Information Philadelphia USA.
3. Egghe L, Rousseau R (1990) Introduction to informetrics. Amsterdam, Elsevier.
4. Börner K, Maru JT, Goldstone RL (2004) The simultaneous evolution of author and paper networks. Proc Natl Acad Sci USA 101: 5266-5273.
5. Figg WD, Dunn L, Liewehr DJ, Steinberg SM, Thurman PW, et al. (2006) Scientific collaboration results in higher citation rates of published articles. Pharmacotherapy 26: 759-767.
6. Garfield E (2004) Historiographic mapping of knowledge domains literature. J Inform Sci 30: 119-145.
7. McCain KW (2008) Assessing an author's influence using time series historiographic mapping: the oeuvre of Conrad Hal Waddington (1905-1975). J Am Soc Information Sci Tech 59: 510-525.
8. Tutarel O (2004) Composition of the editorial boards of leading medical education journals. BMC Med Res Methodol 4: 3.
9. Malin B, Carley K (2007) A longitudinal social network analysis of the editorial boards of medical informatics and bioinformatics journals. J Am Med Inform Assoc 14: 340-348.
10. Newman MEJ (2004) Coauthorship networks and patterns of scientific collaboration. Proc Natl Acad Sci USA 101: 5200-5205.
11. Patel V, Kim YR (2007) Contribution of low- and middle-income countries to research published in leading general psychiatry journals, 2002-2004. Br J Psychiatry 190: 77-78.
12. Patel V, Sumathipala A (2001) International representation in psychiatric literature: survey of six leading journals. Br J Psychiatry 178: 406-409.
13. Sumathipala A, Siribaddana S, Patel V (2004) Under-representation of developing countries in the research literature: ethical issues arising from a survey of five leading medical journals. BMC Med Ethics 5: 5.
14. Mahawar KK, Malviya A, Kumar G (2006) Who publishes in leading general surgical journals? The divide between the developed and developing worlds. Asian J Surg 29: 140-144.
15. Andrews JE (2003) An author co-citation analysis of medical informatics. J Med Libr Assoc 91: 47-56.
16. McCain KW (1990) Mapping authors in intellectual space: a technical overview. J Am Soc Information Sci 41: 433-443.
17. Laband DN, Piette MJ (1994) A citation analysis of the impact of blinded peer review. JAMA 272: 147-149.
18. Eysenbach G (2006a) Citation advantage of open access articles. PLoS Biol 4: e157.
19. Eysenbach G (2006b) The open access advantage. J Med Internet Res 8: e8.
20. Baltussen A, Kindler CH (2004a) Citation classics in anesthetic journals. Anesth Analg 98: 443-451.
21. Baltussen A, Kindler CH (2004b) Citation classics in critical care medicine. Intensive Care Med 30: 902-910.
22. Meneghini R, Packer AL (2006) Articles with authors affiliated to Brazilian institutions published from 1994 to 2003 with 100 or more citations: II - identification of thematic nuclei of excellence in Brazilian science. An Acad Bras Cienc 78: 855-883.
23. Packer AL, Meneghini R (2006) Articles with authors affiliated to Brazilian institutions published from 1994 to 2003 with 100 or more citations: I - the weight of international collaboration and the role of the networks. An Acad Bras Cienc 78: 841-853.
24. Paladugu R, Schein M, Gardezi S, Wise L (2002) One hundred citation classics in general surgical journals. World J Surg 26: 1099-1105.

25. Tsai YL, Lee CC, Chen SC, Yen ZS (2006) Top-cited articles in emergency medicine. Am J Emerg Med 24: 647-654.

26. Merton RK (1973) The sociology of science. University of Chicago Press, London.

27. Price DDS (1976) A general theory of bibliometric and other cumulative advantage processes. J Am Soc Information Sci 27: 292-306.

28. Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286: 509-512.

29. Kleinberg JM, Kumar R, Raghavan P, Rajagopalan S, Tomkins AS (1999) The web as a graph: measurements, models, and methods. Computing and Combinatorics 1627: 1-17.

30. Gami AS, MontoriVM, Wilczynski NL, Haynes RB (2004) Author self-citation in the diabetes literature. CMAJ 170: 1925-1927.

31. Callaham M, Wears RL, Weber E (2002) Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. JAMA 287: 2847-2850.

32. Kaltenborn KF (2004) Commentary III - Validity and fairness of the impact factor – - a comment on the article by Decker et al. Sozial- und Präventivmedizin/ Social and Preventive Medicine 49: 23-24.

33. Scully C, Lodge H (2005) Impact factors and their significance; overrated or misused? Br Dent J 198: 391-393.

34. Horrobin DF (2001) Something Rotten at the Core of Science? Trends in Pharmacological Sciences 22: 51-52.

35. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99: 7821-7826.

36. El-Arini K, Guestrin C (2011) Beyond keyword search: discovering relevant scientific literature. Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, ACM New York, NY, USA.

37. McCain KW (1989) Descriptor and citation retrieval in the medical behavioral science literature: retrieval overlaps and novelty distribution. J Am Soc Information Sci 40: 110-114.

38. Swanson DR (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 30: 7–18.

39. Swanson DR (1988) Migraine and magnesium: Eleven neglected connections. Perspect Biol Med 31: 526–557.

40. Swanson DR, Smalheiser NR (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence 91: 183–203.