Textual data mining

Sally Jo Cunningham
Dept. of Computer Science
University of Waikato
Hamilton, New Zealand
email: sallyjo@waikato.ac.nz

Abstract: Most automated or semi-automated techniques for extracting novel information from data have concentrated on analyzing simple tables of numeric or atomic symbolic values. A related (but much more complex) problem, that of inferring new facts or knowledge from textual databases, has been addressed most effectively by the library and information retrieval research communities. This paper incorporates several ad hoc search strategies proposed by those communities into a single search methodology that guides the search process and provides a framework for the presentation of facts gleaned from the search. This graphical search result representation is semi-formal, in the sense that it represents the *structure* of search results formally while the *contents* of the search are represented informally. The methodology is intended as an aid to "mining" new scientific information from textual/bibliographic databases, rather than as an automated proof system.

I. Introduction

The dramatic growth, both in number and size, of scientific databases has sparked a considerable interest in the field of machine learning and the related study of data mining techniques. As first megabytes, then gigabytes, and now terabytes of data are being collected [PI91], it has become progressively less feasible to produce a timely analysis of the accumulation; hence the development of the fields of *machine learning* (for example, see [QU86], [CE87], [WI93]) and *data mining* (for example, see [[KL92], [PI91]).

These techniques provide automated or semi-automated methods for extracting implicit and previously unknown information from the raw data. The data analysis algorithms are for the most part limited to handling relational-type tables, where each example (tuple) is represented by a fixed length vector of attribute values. Attributes are generally restricted to simple numeric or symbolic values, and attribute vectors cannot describe situations that involve more complex structures or relations between objects. Most machine learning algorithms require carefully hand-crafted tables of examples, and induce a decision tree or classification rule set that provides a model of the application domain. In contrast, data mining techniques attempt to gather less structured information from more realistic databases; they typically explore ("mine") a rich, relatively unstructured set of data and retrieve from it unusual patterns, unexpected regularities, implicit information, etc. These interesting findings are generally not complete rules, but may suggest information that should be incorporated into a rule base.

The far more complex problem of discovering information in textual (often bibliographic) databases has been addressed primarily in the library and information science community, through ad hoc bibliographic search techniques. Most bibliographic searches are carried out with the intention of retrieving information that is new to the searcher but not to the scientific community. In a remarkable series of papers, however, Roy Swanson has demonstrated that novel information can be extracted from a known literature ([SW86], [SW87], [SW89b]). He has discovered previously unnoticed logical connections between unrelated topics in the medical literature--noting in one case that fish oils have

characteristics that may alleviate the symptoms of Raynaud's syndrome, and in another that magnesium deficiency may be a cause of some migraines. Finding these connections between disparate literatures involved noting assertions of the type 'A causes B' and 'B causes C', and inferring the transitive assertion that 'A may cause C'. Both hypotheses have since been confirmed by clinical trials ([DI89], [DA89]).

Swanson also presented a systematic, trial-and-error search strategy for locating what he has christened "undiscovered public knowledge" in bibliographic databases [SW89a]. Other information science researchers have expanded on Swanson's methodology [DA89], or have suggested additional search techniques that may be brought to bear on this problem ([HA84], [PO87], [BA89]). This paper describes a procedure that supports these techniques by incorporating them into a single search methodology, guiding the search process and organizing the presentation of facts gleaned from the search.

The search domain is the MEDLINE medical database. This database was chosen for several reasons:

- medical problems can generally be described in terms of cause and effect, which lends itself to the proposed search procedure and search results display;
- documents in the medical literature have structured, informative titles, which enhances the process of browsing through large numbers of bibliographic entries;
- the database's medical subheadings (MeSH) terms can be used to focus the set of documents retrieved for inspection, reducing the number of irrelevant document references retrieved; and
- most document references in the MEDLINE database contain abstracts, which in many cases sufficiently describe the paper's results so as to eliminate the need to retrieve the paper itself.

This approach may be generalizable to other databases with similar attributes--for example, collections of patent descriptions or legal texts. Indeed, the representation for search results is heavily based upon experimental systems for graphically describing case supports for legal arguments ([RO90], [HO94]]).

This paper is organized as follows: Section 2 describes the search structure for "mining" a medical bibliographic database; Section 3 presents a graphic representation for organizing and summarizing the results of the search; and Section 4 discusses areas for further research.

II. Search structure

The search is conducted through a series of stages: exploratory browsing through the literature to determine symptoms or disease traits that merit exploration; investigation of these symptoms outside the context of the original disease, to locate relevant information not previously connected with the disease under study; confirmation of the novelty of the connections inferred; and formal representation of the argument relating the disease to its cure. These stages may overlap, as, for example, the searcher leaps from an initial search to a potential cure, or as a particularly promising argument is described during the intermediate search state.

Stage 1: exploratory

The searcher begins by browsing a list of article titles that contain the name of the disease or condition under study. The MeSH subheadings appropriate to that disease are displayed, together with the number of documents in the database indexed under the disease/subheading pair. The document sets may be selected for browsing, and promising terms from the titles, abstracts, or descriptors are entered into a pool of search terms for Stage 2.

The MEDLINE interface supports this process by providing access to displays of the tree structure of the MeSH subheadings, greatly increasing the ease of browsing. Subheadings can be used to more narrowly focus or to generalize the search:

Narrowing the search space: for example, in Swanson [SW87] the set of titles containing the term "Raynaud" (for Raynaud's syndrome) was limited by several MeSH terms, among them the subheading "Blood" (indicating that the disease causes changes in the blood). Browsing these titles and abstracts led to the identification of several recurring terms such as "blood viscosity" and "red cell deformability". Note that the MeSH subheadings can be extremely effective in organizing and limiting searches to items most likely to be relevant; for example, a keyword search on "blood" would lead to an overwhelming torrent of (predominantly irrelevant) items. Moreover, there are generally relatively few subheadings to explore--as Swanson also notes, "only 42 are applicable to diseases, at all, and many of these 42 would not be reasonable choices for [an] exploratory search for a cure" [SW87] (eg, the "economics", "legislation and jurisprudence", and "manpower" subheadings).

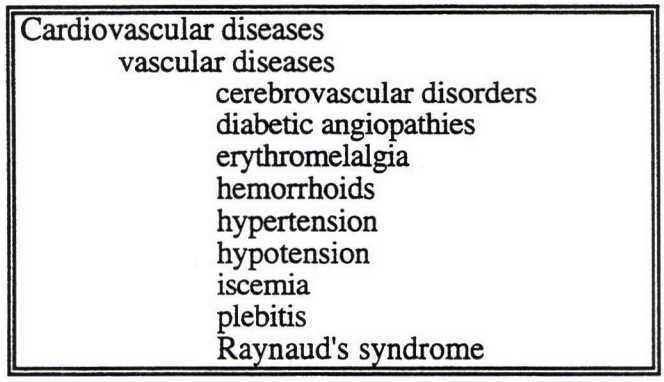

Figure 1. Subset of the MeSH headings related to Raynaud's syndrome

Generalizing to produce analogies: Davies suggests a simple mechanism for searching for novel solutions to apply to problems: generalize the specific problem under study, and then examine the solutions to any closely analogous problems revealed by the generalization [DA89]. In MEDLINE, travelling "up" the MeSH tree yields related diseases and general categorizations of disorders. For example, Figure 1 presents a subset of the tree related to Raynaud's syndrome. Treatments for these associated diseases and disorders may suggest treatment for the specific disease under study.

Stage 2: extension

The set of terms selected in Stage 1 are searched, and titles retrieved are browsed to locate factors affecting them. In the Raynaud's syndrome example, examination of the blood viscosity and deformability literature led to a pair of articles indicating that fish oil increases

red blood cell deformability and decreases blood viscosity. Since Raynaud's disease is a peripheral circulatory disorder, these articles appeared promising--the flow of blood in small blood vessels is improved by a higher degree of blood cell deformability and increased fluidity.

If the search sets are initially too large to browse, limiting terms can be selected from a ranked display of frequently occurring descriptors and terms extracted from the set. This is a common technique for "cherry-picking" from large search sets ([BA89], [DA89]).

Stage 3: confirmation of novelty

In this type of database exploration, the search emphasis is on creating new hypotheses from the literature, rather than the more conventional goal of locating relevant existing conclusions. At this point, then, the novelty of the hypothesis derived in Stage 2–in the example, that fish oil can favorably affect Raynaud's syndrome–is tested by a literature search (here, for articles on both fish oil and Raynaud's). If the search retrieves articles indicating such a link, then the hypothesis has already been proven/disproven, and the mining process begins again in Stage 1. If no document is located that connects the two literatures, then the hypothesis is further explored for implicit connections between the two areas medical study, and the connections located are analyzed for their plausibility (Stage 4).

At this point Swanson advocates performing co-citation analysis on the two bodies of literature–here, checking whether the Raynaud's syndrome papers and the papers describing the effects of fish oil on the blood share any references. By examining both one-step links between documents (direct citations of each other) and two-step links (co-citations), Swanson ensures that the connections between the ideas in the two sets of papers are completely unexplored. On the other hand, if the hypothesis has been so thoroughly overlooked or ignored that it appears only implicitly through co-citation analysis, then it is still relatively novel and suitable for further analysis. Given that co-citation analysis is also a relatively time-consuming process, it appears reasonable to abandon its use in this context.

Stage 4: Constructing a representation of supports for the hypothesis

The links between disease, symptoms, possible cures, and supporting documents are formally represented in a graphical notation. This notation is similar to the semantic net-like schemes used to represent case supports for legal arguments [RO90]. As noted above, the argument may be constructed simultaneously with the extension phase of the search. The format for this notation is presented in Section 3 below.

A graphic representation of the search results is helpful in that the transitive "A causes B", "B causes C", and therefore "A causes C" relationships Swanson proposes are not often that clear-cut. Indeed, in his examples the supports for his hypotheses are held to varying degrees: "A is a sole cause/ a major cause / tends to cause / is one cause among many/ of B" [DA89]. Sources may provide contradictory evidence, or may provide weak rather than strong support for the hypothesis (if additional explanations of the same results are possible). Swanson's support for the connection between magnesium deficiency and migraines, for example, is based on a discussion of 11 weak supports that, taken together, provide a convincing argument for the hypothesis. Organizing the mining results into a coherent form could be useful in directing the mining and in evaluating the plausibility of the hypothesis.

III. Representation of hypothesis support

The evidence supporting the hypothesis being "mined" can be represented by a notation based on those used to graphically present legal arguments ([RO90], [HO94]). While far more complex notations exist for structuring medical knowledge [PO87], the emphasis in this work is on a less detailed, but correspondingly easier to use, technique. The anticipated users would create the graphic structure to organize a search, rather than as a basis for further automated inference, and hence would both require and desire a simpler, less rigorous notation.

Each assertion or argument used to support or refute the hypothesis is represented by a node in a tree, arranged in a top-down hierarchy from the most general (the hypothesis) to the most specific (a scientific fact or a reference to a document). Links between the nodes represent the node's effect in supporting or weakening the hypothesis. The notation includes the following symbols (Figure 2):

- hypothesis: major statement to be supported or refuted
- argument : a statement drawn from the literature that serves to support or rebut the main hypothesis or another argument
- fact: a generally accepted statement that does not require additional support
- document node: indicates an individual reference
- supporting or rebutting links: indicating the relationship between argument, fact, and hypothesis nodes

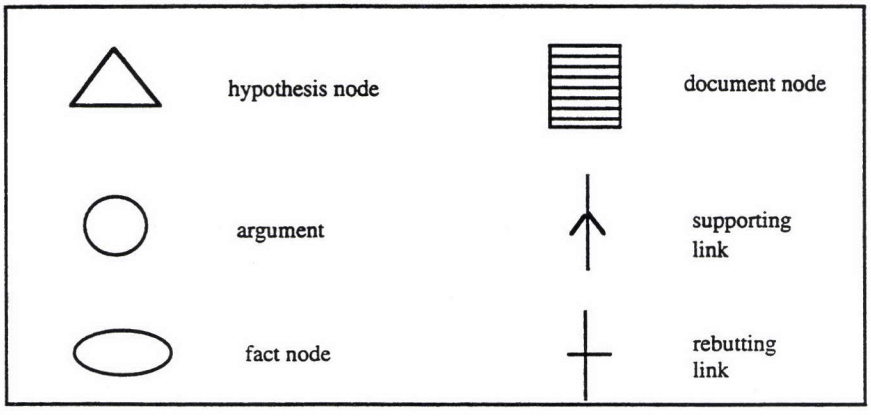

Figure 2. Graphical notation for representing mining results

So, for example, a small portion of the results of Swanson's mining of MEDLINE for potential cures for Raynaud's syndrome could be represented as:

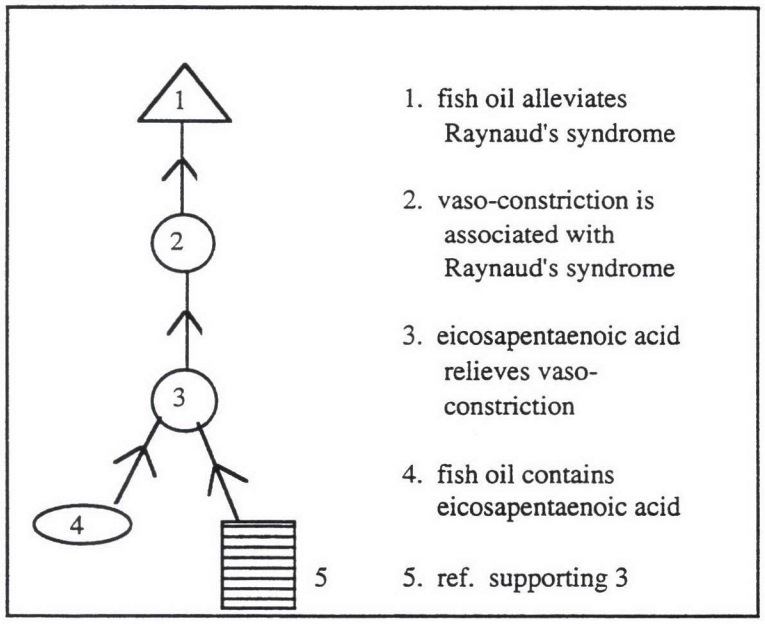

Figure 3. Portion of support for the hypothesis on fish oil and Raynaud's syndrome

IV. Conclusions

The search process is perhaps best described as guided serendipity: the above described techniques and tools provide direction in locating the most promising search paths and support the rapid elimination of mistaken guesses. This "mining" methodology is made more attractive by the widespread availability of MEDLINE on CD-ROM, since exhaustive, open-ended searches can now be conducted cheaply. And, of course, bibliographic database mining is extremely cost-effective in that it can be used to guide clinical or laboratory research into more promising paths--illustrating the truth of the old librarian's bromide that "six months in the lab can save you an afternoon in the library"!

This approach to detecting and representing new medical hypotheses that are created by combining existing information in novel ways can be implemented over existing bibliographic databases (in this case, MEDLINE). Other techniques for inferring novel information suggested in the literature [DA89] have the disadvantage of requiring extensive re-indexing of the databases (for example, using Farradane's relational indexing technique [FA73]). While these techniques have the advantage of promising a more automated approach to generating new information from textual databases, the high costs of providing the initial semantic analysis and re-indexing appear too great--at least until an automated semantic indexing scheme is provided!

The natural final step beyond the proposed user-guided bibliographic data "mining" process would be to automatically extract facts from the document abstracts and titles, and use these facts as a basis for automated inference. Systems do exist that can locate and manipulate facts in limited styles of text; for example, JASPER (Journalist's Assistant for Preparing Earnings Reports) is a fielded system that processes live feed of company press releases from PR Newswire to produce first drafts of Reuters news stories [HA92]. The cost of building such an application is high, however--for JASPER, estimated at eight person-months of knowledge engineering for its relatively narrow domain. As the point of database mining is to permit the combination of information from previously disjoint literatures, it would be difficult to use JASPER's approach on specialized subsets of a bibliographic database. Additionally, the sheer volume of facts extractable from a database

like MEDLINE, in combination with the mass of supporting "common sense" information from the medical literature, would create a search space that would appear to require expert intervention to efficiently mine.

References

[BA89] Bates, M.J. (1989): "The design of browsing and berrypicking techniques for the online search interface". *Online Review* no. 5, pp. 407-424.

[CE87] Cendrowska, J., 1987: "PRISM: an algorithm for inducing modular rules." *Int J Man-Machine Studies 27* (4), pp. 349–370.

[DI89] DiGiacomo, R.A., Kremer, J.M., and Shah, D.M. (1989): "Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study". *American Journal of Medicine 86* (2), pp. 158-164.

[DA89] Davies, R. (1989): "The creation of new knowledge by information retrieval and classification". *Journal of Documentation 45* (4), pp. 273-301.

[FA73] Farradane, J., Russell, J.M., and Yates-Mercer, P.A. (1973): "Problems in information retrieval: logical jumps in the expression of information". *Information Storage and Retrieval 9* (2), pp. 65-77.

[HA84] Harter, S.P. (1984): "Scientific inquiry: A model for online searching". *Journal of the American Society for Information Science 35* (2), pp. 110-117.

[HA92] Hayes, P. (1992): "Intelligent high-volume text processing using shallow, domain-specific techniques". In *Text-based Intelligent Systems: Current research and practice in information extraction and retrieval*, ed. by P.S. Jacobs. New Jersey: Lawrence Erlbaum Associates.

[HO94] Hosking, P. (1994): *Argument representation and conceptual retrieval for litigation support.* Master's Thesis, Victoria University of Wellington, New Zealand, Dept of Computer Science.

[KL92] Klosgen, Willi (1992). "Problems for knowledge discovery in databases and their treatment in the statistics interpreter EXPLORA". *Int J for Intelligent Systems 7* (7), pp. 649-673.

[PI91] Piatetsky-Shapiro, G., and Frawley, W.J., editors (1991): *Knowledge discovery in databases.* Menlo Park, CA: AAAI press/The MIT Press.

[PO87] Pollitt, S. (1987): "CANSEARCH: An expert systems approach to document retrieval". *Information Processing and Management 23* (2), pp. 119-138.

[QU86] Quinlan, J.R., 1986: "Induction of decision trees." *Machine Learning 1* (1), pp. 81–106.

[RO90] Robertson, B. "John Henry Wigmore and Arthur Allan Thomas--An example of Wigmorian Analysis". *VUW Law Review 20*, Victoria University of Wellington.

[SW86] Swanson, D.R. (1986): "Undiscovered public knowledge". *Library Quarterly 56* (2), pp. 10-3-118.

[SW87] Swanson, D.R. (1987): "Two medical literatures that are logically but not bibliographically connected". *Journal of the American Society for Information Science 38* (4), pp. 228-233.

[SW89a] Swanson, D.R. (1989): "Online search for logically-related noninteractive medical literatures: A systematic trial-and-error strategy". *Journal of the American Society for Information Science 40* (5), pp. 356-358.

[SW89b] Swanson, D.R. (1989): "A second example of mutually isolated medical literatures related by implicit, unnoticed connections." *J of the Am Society for Information Science 40* (6), pp. 432-435.

[WI93] Witten, I.H., Cunningham, S.J., Holmes, G., McQueen, R., and Smith, L. (1993): "Practical Machine Learning and its Application to Problems in Agriculture". *13th Conference of the New Zealand Computer Society*, vol. 1, Auckland, New Zealand, August 1993, pp. 308-325.