MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

An Information Extraction and Representation System for Rapid Review of the Biomedical Literature

Debra Revere[a], Sherrilynne Fuller[a], Paul F. Bugni[a], George M. Martin[b]

[a] *Telemakus Research Program, Division of Biomedical & Health Informatics, University of Washington, Seattle, WA, USA*
[b] *Department of Pathology, School of Medicine, University of Washington, Seattle, WA, USA*

Abstract

With the rapid expansion of scientific research, the ability to effectively find or integrate new domain knowledge in the sciences is proving increasingly difficult. The development of methods and tools for assisting researchers to effectively extract problem-oriented knowledge from heterogeneous and massive information sources, and for using this knowledge in problem-solving is one of the most fundamental research directions for the information and computer sciences today. There is a need for new tools to support more precise identification of relevant research articles and provide visual clues regarding relationships among the document sets. We present the Telemakus system in which aggregated citation information and extracted research findings are displayed in a schema-based document surrogate and an interactive mapping tool provides graphical displays of research inter-relationships from documents across a domain. This system is an innovative approach to creating useful and precise document surrogates and may re-conceptualize the way we currently represent, retrieve, and assimilate research findings from the published literature.

Keywords:

Databases, Bibliographic; Databases, Factual; Information Storage and Retrieval; Information Systems; Unified Medical Language System

Introduction

Biomedical researchers are increasingly in need of information systems that actually extract desired facts and answer questions. A researcher will often have an information need that could be expressed as, "Has anyone studied the relationship between *concept A* and *concept B* (e.g., caloric restriction and cancer)? If so, what type of animal was used, what type of experiments were done and what were the findings?" A successful response to a query of this type is extremely difficult in traditional information retrieval (IR) systems. Typically, IR systems operate using a standard query-retrieval model, in which a user describes a need for information with a set of query words and the IR system supplies a list of potentially relevant documents within the searched collection. The returned document list is usually the result of a comparison of the user's query words to certain selected fragments of the original document—e.g., title, abstract, keywords—that are presumed to represent a document's content.

These fragments, however, do not necessarily serve as either summaries or surrogates for document content. Titles do not completely convey content and abstracts, though indicative of content, are not fully informative. The problems of indexing [1], indexer-indexer inconsistency [2] and capturing the "aboutness" of a document [3] have been discussed extensively in the literature. Biomedical researchers need a system that will produce a coherent, but brief, document representation that provides a "window" into the original document. They do not have the time to retrieve whole documents that must be read through to find the (potentially) desired information. Also, as the numbers of biomedical publications increase, so do the number of documents that match a user's query; as the retrieval list lengthens, the ability to effectively find or integrate new domain knowledge in the sciences can become increasingly difficult.

In addition to these issues, biomedical researchers are in need of retrieval systems that will help them synthesize information extracted from multiple documents to provide an overview of a subject or to identify new relationships between facts and synthesize new knowledge [4]. When reviewing the research literature, researchers typically focus first on the research findings as reported in the data tables and figures—a hallmark of all reports of original scientific research. However, none of the current major bibliographic databases provide access to the content of the legends from the tables and figures, either directly by listing them or indirectly by facilitating searches of keywords in the legends. A system that allows researchers to quickly review retrieved results for research methods and findings or to quickly view the relationships *among* the documents in the document set would enable scientists to keep abreast of published findings in their direct area of interest, as well as the crossover topics of importance to their field.

It may seem obvious, then, that representing documents by summarizing their content would enable both a reduction in user time needed to review a lengthy retrieval list and more efficiently focus the user's query so the list will contain fewer documents for scientists to read. Numerous approaches to the challenge of producing document surrogates have been explored, including linguistic approaches, such as discourse structure [5], lexical cohesion [6] and lexical chains [7]; statistical approaches, such as maximal marginal relevance [8]; or combinations of the two [9]. All vary in deciding which components of the text are most useful, what level of granularity is most representative of content and how the document summary interface is represented.

In an aptly titled report, "Discovering the information that is lost in our databases: Why bother storing data if you can't find the information?" Bruza and Proper emphasize that providing smooth access paths for retrieving stored information is at least as important as ensuring integrity of the stored information [10]. Standard access approaches, discussed above, fall short of providing access to the data *within* the content.

Here we present an innovative system in development at the University of Washington for accessing the data within research documents. Telemakus (http://www.telemakus.net/) is unique in combining document surrogates with interactive concept maps of linked relationships across groups of research reports. Document surrogates present aggregated citation information, research methods and research findings in a conceptual schema interface, enabling users to rapidly access and review biomedical literature.

Background

The unique strength of the Telemakus system lies in the combination of document surrogates with interactive maps of linked relationships across groups of research reports. The system can be divided into three integrated components to retrieve, display and summarize biomedical research reports across a domain: 1) a Research Report Schema containing research methods and findings extracted from original documents presented in a consistent, coherent and structured schema format that functions as a document; 2) Research Concept and Relationship Extraction incorporating controlled vocabulary to index the research findings extracted from data tables and figures; and 3) a Visual Exploration Interface providing a dynamic, graphical map of research findings. Each component is detailed below.

Research Report Schema as Document Surrogates

Research papers have a highly predictable structure [11,12], including typical structural elements such as "abstract", "background", "purpose", "methods", "materials", "results" and "discussion." The Telemakus information extraction (IE) system leverages this predictable structure of research reports and analyzes the full-text document in order to extract pre-defined elements within the Materials, Methods and Results (including the tables and figures) sections.

The document surrogate includes: standard bibliographic information (author, title, journal), research design and methods information (age, sex, number of subjects, pre-treatment and treatment regimen, organism and source of organism) and research findings derived from data tables and figures. The layout of the document surrogate schema or template is based on the application of schema theory to scientific research (e.g., schematic representation of psychological reports [13], clinical trials [14] and superstructure and predictability of text [15]). The Telemakus system has extended the schema to creating surrogates of biological research reports with representations of research methods and findings [16,17].

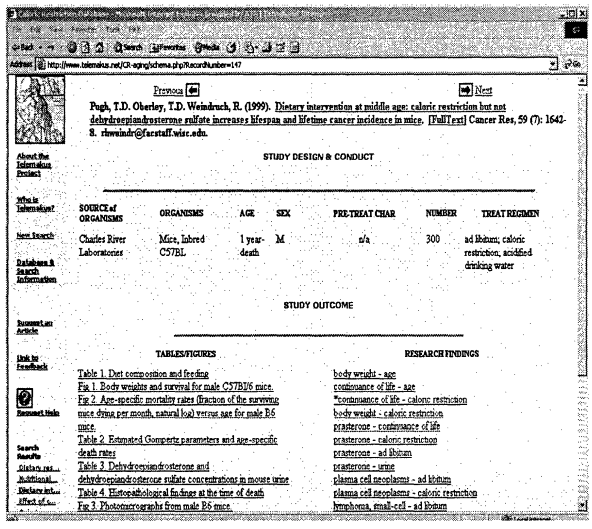

Figure 1 - Populated schema for a retrieved document

A schema-based document template is populated with data extracted for each field, as in Fig. 1. The schema provides a consistent roadmap for users to read and browse through numerous research reports. The interface serves as a search tool as well as provides links to full-text (if available) and the actual tables and figures. Research findings can be searched directly from the schema, offering a rapid way of following research connections through the database. It is also possible to search on other IE elements (abstract, keywords) that are stored in the database but not displayed in the schema.

Research Concept and Relationship Extraction

A document's tables and figures are the one place where researchers unambiguously present their research findings. Researchers often initially focus on the data found within tables and figures (sometimes before or instead of reading the article). Extracting the headings and providing linked research concepts mimics a researcher's traditional approach to reading the research literature [18]. Telemakus uses research concepts and their relationships—typically linked together as pairs (the *x-y* axes of the tables and graphs)—to represent a document's contents, providing a means of reviewing concepts studied both *within* and *across* a set of retrieved documents.

We use the Unified Medical Language System Metathesaurus (UMLS META) as the basis for controlling concept vocabulary throughout the database. META is a database of information on concepts that appear in one or more of a number of different controlled vocabularies and classifications, providing a uniform, integrated distribution format from over 95 biomedical vocabularies and classifications [19]. A research concept identified in a document's data tables and figures is mapped to its META preferred term, along with its synonyms, semantic type (e.g., Disease or Syndrome), broader and narrower terms and

Unique Identifier (used for future automated updates of the thesaurus). An example of a source figure and its extracted concepts and relationships is shown in Fig. 2.

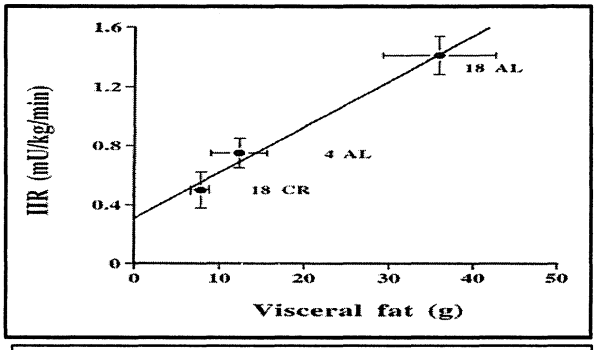

Figure Heading: "The relationship between insulin infusion rate (IIR) and visceral fat (VF). Points presenting the 4 ad libitum (AL), 18 AL, and 18 caloric restricted (CR) rats."
Extracted Research Concepts & Relationships:
insulin infusion rate - visceral fat
insulin infusion rate - ad libitum
insulin infusion rate - caloric restriction
visceral fat - ad libitum
visceral fat - caloric restriction

Figure 2 - Research concept and relationship extraction

Visual Exploration Interface

There is a growing body of work (e.g., [20,21]) related to mapping metaphors and visualizing large document sets and database search results to provide the user with the ability to visualize relationships among documents and their contents. In addition, several tools have been developed that graphically present inter-document relationships, most commonly using some form of link-node diagram [22].

Concept mapping is a technique for representing knowledge graphically through networks of concepts. Such networks consist of nodes (points) and links (arcs/edges). Nodes represent concepts and links represent connections between concepts. Concept mapping has been used for a variety of purposes, such as designing a complex structure, communicating complicated ideas and, in the Telemakus system, to demonstrate connections among ideas.

While the schema acts as a surrogate for individual documents in a retrieval set, the Visual Exploration Interface provides a dynamic map of research findings of the relationships between research concepts across a set of documents. Users can "browse" documents to answer a question such as "Has a statistically significant relationship ever been reported in any animal model between level of fat intake and cancer?" by exploring the relationship "dietary fats – neoplasms." Fig. 3 is a close-up of the Telemakus visual exploration interface. Following any of the research relationship links (e.g., antioxidants – neoplasms) will search and display additional research report schemas with that research finding.

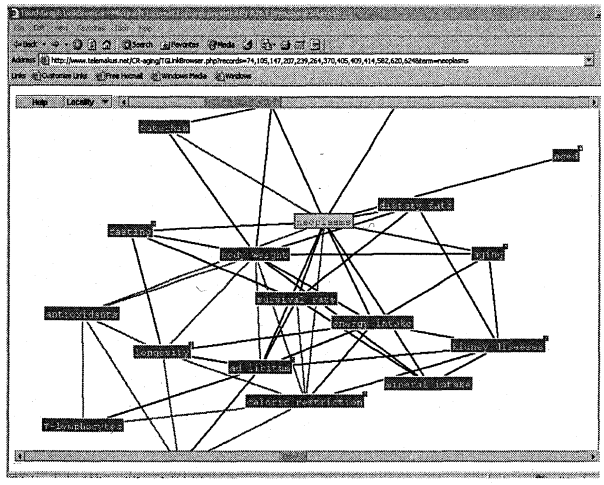

Figure 3 - Map of "neoplasms" research findings

Methods

The Telemakus IE system architecture centers on a relational database, a set of tools used to populate the database with data extracted from research reports and several server side tools and programs responsible for serving the content of the database to the public via the WWW. The entire system is built from open-source components, leveraging standard protocols and tools whenever possible. By design, no platform or language-specific dependencies exist between system components. The goal is to keep each component decoupled, so that any phase may be enhanced without requiring significant changes from other components. Also, this independence makes it possible to add additional components or implement remote processing very efficiently. Fig. 4 details the system architecture.

The IE system currently consists of three discrete phases: Fetcher, Extractor and CrossCheck, each independently responsible for its own task. Fetcher gathers all the resources necessary to process a given research report. It takes the document URL and citation information, creates an XML document and unique directory on the server and downloads the full-text document and its table/figure files. Fetcher then creates local copies of the HTML and image files and creates elements in the XML document for each of the respective pieces.

The XML document is then passed on to Extractor which extracts data from the resources made available by Fetcher. Working with the XML document, whose elements define where the report tables and figures are stored on the server, Extractor identifies sentence boundaries and report sections and creates a list of proposed database entries. Meta tags are inserted in the local copies of the report as reference points to all the extracted data. Elements are added to the XML document, defining all the extracted database fields and their respective reference tags. The XML document is then passed on to CrossCheck which is responsible for presenting the extracted data from the previous

phases to a human analyst for review. CrossCheck provides a visual display of the extracted database fields, selectively highlighted. The analysts can then add, delete or modify the extracted data before committing the report to the database.

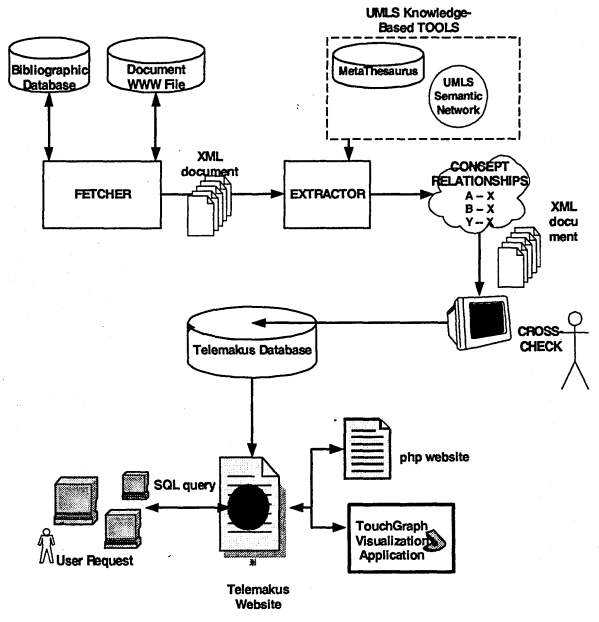

Figure 4 - Telemakus system architecture

An open-source Java-based graphing applet, TouchGraph (http://www.touchgraph.com), has been adapted and enhanced to create dynamic maps of research findings based on the research concept queried either via the schema or retrieval set. The visualization package serializes maps to and from XML. Servlet code delivers an XML document to TouchGraph via HTTP, for the relationships defined in the retrieval set. The package supports dynamic content feeds to generate interactive nodes-and-edges maps. The visualization tool permits traversal from node to node and expanding or contracting the view to include a map of all research relationships reported in the retrieved set of documents and to easily return to query-paired terms of interest. Tool bars permit narrowing and broadening of the focus and rotation of maps for improved viewing.

Telemakus, then, provides two views for the user: the document surrogate and the interactive visualization interface. Using multiple views allows the user to explore concept relationships that represent a domain's research findings in aggregate while retaining the capability to drill-down to an individual report's attribute level.

The initial Telemakus database domain is the biology of aging, although Telemakus tools and concept mapping algorithms are broadly applicable to any domain that presents research findings as numeric data (i.e., in tables and/or figures). The current database and visualization model represents knowledge related to caloric restriction and aging, a subset within the domain of the biology of aging chosen because it is an important rapidly expanding specialized area of the biology of aging that is also highly interdisciplinary.

Results

A primary goal of Telemakus is to create a user-centered product so we have involved researchers in the iterative design and testing of the system. Informal usability indicates that the schematic document surrogate is a major improvement over the traditional bibliographic citation format with abstract because it provides a more comprehensive retrieval set picture for the user. Feedback also affirms that retrieval based on research findings is a unique and highly desirable core function. Users have also expressed the need for more control over both the schema and mapping presentation in order to fine-tune the display to individual preferences and personal needs (e.g. color blindness). Also, as more concepts are present the display becomes increasingly crowded (a "hairball" effect); users need a variety of ways to prune the number of concepts so the presentation remains meaningful.

Discussion

From its inception, Telemakus has been designed in close collaboration with the scientists who are its users. We plan to continue our iterative design and test approach to improve and enhance both IE and presentation with the following evaluations: (1) value vs. cost of human review and (2) map filtering experiments based on a variety of criteria, including Semantic Type or organism studied (e.g., human versus rat) to address "hairball" and other display issues. In addition, we are planning a comprehensive evaluation that includes both intrinsic (testing of the system in of itself) and extrinsic (testing how use of the system affects the completion of other tasks) evaluations.

Conclusion

As the compiled record of scholarly knowledge has grown exponentially, it has become impossible to remain abreast of all relevant scientific findings. Even with remarkably useful online bibliographic databases, search results continue to overwhelm researchers. It is critical that information tools be developed to address these information overload problems in order to reduce redundant research as well as to ensure that scientists can put disparate findings together to develop new research hypotheses. The Telemakus system provides a flexible new method for rapid review of documents. By formalizing representation of the methods and results of scientific research reports, Telemakus offers the potential to ultimately speed up the scientific discovery process.

Acknowledgments

We wish to thank Heather Fuller, Wendy Kramer, David Owens, Lucas Reber and Lisa Tisch for their essential contributions to this research effort. We gratefully acknowledge funding provided by the Ellison Medical Foundation.

References

[1] Bates MJ. Subject access in online catalogs: A design model. *J Am Soc Inf Sci*, 1986;37:357-76.

[2] Chung Y, Pottenger WM, Schatz BR. Automatic subject indexing using an associative neural network. In: *Proc 3rd ACM Conf on Digital Libraries*, 1998; pp. 59-68.

[3] Bruza PD, Song DW, Wong KF. Aboutness from a commonsense perspective. *J Am Soc Inf Sci*, 2000;51:1090-1105.

[4] Khoo C, Myaeng SH. Identifying semantic relations in text for information retrieval and information extraction. In: Green R, Bean CA & Myaeng SH, eds. *The Semantics of Relationships: An Interdisciplinary Perspective.* Boston: Kluwer, 2002; pp. 161-80.

[5] Hahn U, Strube M. Centered segmentation: scaling up the centering model to global discourse structure. In: *Proc 19th Conf of the Cognitive Science Society*, 1997; pp. 104-11.

[6] van Gils B, Paijmans H. Creating document surrogates with lexical cohesion. In: *Proc 3rd Dutch-Belgian Info Retrieval Wkshp*, 2002; pp. 64-68.

[7] Barzilay R, Elhadad M. Using lexical chains for text summarization. In: *Proc ACL'97 Wkshp on Intelligent, Scalable Text Summarization*, 1997; pp. 10-17.

[8] Carbonell J, Goldstein J. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In: *Proc 21st ACM/SIGIR Conf on R & D in IR*, 1998; pp. 335-36.

[9] Strzalkowski T, Stein G, Wang J, Wise B. A Robust Practical Text Summarizer. In: Mani I & Maybury MT, eds. *Advances in Automated Text Summarization.* Cambridge: MIT Press, 1999; pp. 137-54.

[10]Bruza PD, Proper HA. Discovering the Information that is lost in our Databases: Why bother storing data if you can't find the information? Technical report, Distributed Systems Technology Centre, Brisbane, Australia, 1996.

[11]Kando N. Text-level structure of research papers: Implications for text-based information processing systems. In: *Proc 19th BCS-IRSG Colloquium on IR.* London: Springer-Verlag, 1997; pp. 68-81.

[12]Paice CD, Jones PA. The identification of important concepts in highly structured technical papers. In: *Proc 16th ACM SIGIR Conf on Research and Development in IR*, 1993; pp. 69-78.

[13]Kintsch W, van Dijk TA. Toward a model of text comprehension and production. *Psychol Rev*, 1978;85:363-94.

[14]Fuller S. *Schema theory in the representation and analysis of text.* Dissertation, Library Science. Univ Southern California, 1984. 189p. U.M.I. Order No.DA8500206.

[15]Dillon A, Schaap, D. Expertise and the perception of shape in information. *J Am Soc Inf Sci*, 1996;47:786-88.

[16]Fuller S, Revere D, Bugni P, Martin GM. Telemakus: A schema-based information system to promote scientific discovery. *J Am Soc Inf Sci & Tech*, 2004;in press.

[17]Revere D, Fuller S, Bugni P, Martin GM. A new system to support knowledge discovery. In: *Proc of Am Soc Info Sci & Tech*, 2003; pp. 52-58.

[18]Bishop AP. Document structure and digital libraries: how researchers mobilize information in journal articles. *Inform Proc Manage*, 1999;35:255-79.

[19]NLM. *2003 UMLS Knowledge Sources Documentation.* 01/01/2003. Retrieved from: http://umlsks5.nlm.nih.gov/kss/background/umlsReleases/2003AA/DOC/index.html

[20]Chen C. Visualization of knowledge structures. In: Chang SK, ed. *Handbook of Software Engineering and Knowledge Engineering, Vol. II.* River Edge, NJ: World Scientific Pub, 2002; pp. 201-38.

[21]Hetzler B, Harris WM, Havre S, and Whitney P. Visualizing the full spectrum of document relationships. In: *Proc ISKO Conf*, 1998; pp. 168-75.

[22]Wise JA, Thomas JJ, Pennock K, Lantrip D, Pottier M, Schur A, and Crow V. Visualizing the nonvisual: Spatial analysis and interaction with information from text documents. In: *Proc IEEE Symp on Info Vis*, 1995; pp. 51-58.

Address for correspondence

Debra Revere, Box 357155, University of Washington, Seattle, WA 98195-7155, USA; 1.206.221.2992; *drevere@u.washington.edu*