

A New System to Support Knowledge Discovery: Telemakus

Debra Revere*, MLIS, MA

Telemakus Research Program, Box 357155, Dept of Medical Education & Biomedical Informatics, University of Washington, Seattle, WA 98195-7155. drevere@u.washington.edu [*corresponding author]

Sherrilynne S. Fuller^a, PhD, Paul F. Bugni^a, BS, George M. Martin^b, MD

^aTelemakus Research Program, Box 357155, Dept of Medical Education & Biomedical Informatics, University of Washington, Seattle, WA 98195-7155. {sfuller, pbugni}@u.washington.edu

^bDept of Pathology, Box 357470, School of Medicine, University of Washington, Seattle, WA 98195-7470. gmmartin@u.washington.edu

An unfortunate consequence of specialization in the sciences is poor communication across domains—which can hamper the knowledge discovery process. Research findings in one area may be pertinent to another, researchers may be unaware of relevant work by others that could be integrated into theirs, and important findings just outside a researcher's focus may go undiscovered. Compounding this problem is the information overload issue—the difficulty of keeping current with information that continues to grow at an exponential rate. The development of methods and tools for assisting researchers and other professionals in an effective extraction of 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. It is clear that there is a need for new tools to support more precise identification of relevant research articles and, further, to provide visual clues regarding relationships among the document sets. We present here a suite of such tools which has been in development at the University of Washington for several years.

Introduction

The goal of information retrieval systems is to identify the most precise or relevant subset of documents in a collection that are likely to contain relevant information in response to a user's query by matching concepts (typically keywords) in documents with concepts in the query statement. This is, in fact, what PubMed and other bibliographical retrieval systems do admirably. But biomedical researchers increasingly are seeking systems that actually extract desired facts and answer questions, not

just systems that retrieve whole documents that must be read through to find the (potentially) desired information. And scientists 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 (Khoo & Mayeng, 2002). They need a system that allows them to quickly review retrieved results for research methods and findings or to quickly view the relationships among the documents in the document set.

In addition, 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. Queries such as "has anyone studied the relationship between *concept A* and *concept B* (e.g., *caloric restriction* and *cancer*)" cannot be answered by current information retrieval systems. Although several information systems have been developed to promote scientific discovery, much of this work is based on traditional search and retrieval approaches and the tools for locating and inter-relating research methods and findings are very imprecise in spite of great improvements in the automation of document retrieval over the past twenty years (Friedman et al., 2001; Piniewski-Bond et al., 2001; Swanson, 2001; Swanson & Smalheiser, 1997).

The Telemakus System (<http://www.telemakus.net/CR-aging/>) is a comprehensive approach to this set of information retrieval and management challenges. In close collaboration with researchers in the basic biology of aging, a working system has been designed to present aggregated citation information, research methods and research findings for display in a conceptual schema, and

Appears in:

Proceedings of the American Society for Information Science & Technology Annual Meeting (pp. 52-58), 10/2003.
copyright ©2003 Information Today.

dynamic visualization tools which present relationships among research findings for multiple documents. Here we describe the theories underlying the Telemakus System (concept representation, schema theory, and information visualization) and an overview of a working implementation designed to enhance the knowledge discovery process through retrieval, visual and interaction tools to mine and map author-reported research findings and their related methods. Initial evaluation results are described and future directions are explored.

Theoretical Basis for Telemakus

The Telemakus system builds on the areas of: 1) concept representation, 2) schema theory, and 3) information visualization to enhance knowledge discovery from the scientific literature.

Concept Representation and Relationships

Concept representation is an important component for accurately representing facts from the document and their interrelationships. Although indexing a document using controlled vocabulary is a standard approach for representing a document, indexing falls short of true document representation because the process reduces the words found in the document to uni-variate and isolated index terms which may be suggestive of the content but are not truly representative of the reported methods and research findings.

Indexing not only obscures the unique organization of the information elements in the document but the relationships among the concepts and their assigned index terms are not necessarily congruent. As Wong et al. (2001) stated, "The information retrieval (IR) problem can be described as a quest to find the set of relevant information objects (i.e., documents D) corresponding to a given information need, represented by a query Q. The assumption is that the query Q is a good description of the information need N. An often used premise in IR is the following: if a given document D is about the request Q, then there is a high likelihood that D will be relevant with respect to the associated information need. Thus the information retrieval problem is reduced to deciding the aboutness relation between documents and queries" (p. 338).

Thus, the characterization and utilization of location of concepts in a scientific document can greatly facilitate accurate document representation. This is the approach of the Telemakus System.

Schema Theory

Schemas are generalized mental models that provide a guide for structuring the process of production and comprehension of texts: "At the simplest level, a schema is a description of a complex object, situation, process or structure. It is a collection of knowledge related to the concept ... providing a guide for structuring the processes

of production and comprehension. In the process of production, a schema ... lists the different parts and properties of a structure which must be decided upon in order to produce it. In comprehension, the set of stored schemas is actively used in a process of "pattern recognition" (Winograd, 1977, pp. 72-74). A classic example of this is the restaurant scenario: if we walk into a restaurant with white tablecloths and napkin-wrapped silverware we will assume it is a more expensive restaurant than the establishment next door that has a napkin dispenser on an uncovered table.

According to schema theory, we understand the world in terms of prototypical patterns: people capture global coherence or structure their knowledge of the world based on scripts, schemas, and narratives in which are embedded a vast array of relationships, concepts, and vocabulary. Schema theory has important implications for the design of information representation and retrieval systems. Research on the application of schema theory to scientific research includes the schematic representation of psychological reports (Kintsch & van Dijk, 1978), clinical trials (Fuller, 1983), and recently, the Telemakus System's application of the research report schema to creating surrogates of reports of basic biological research with representations of research methods and findings (Fuller et al., 2002).

Visual Mapping of Inter- and Intra-Document Relationships

As stated previously, a researcher will often have an information need that could be expressed as, "Has anyone studied the relationship between *concept A* and *concept B*? 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 or impossible in traditional information retrieval systems because "...conventional IR systems that employ isolated term assignments seem inadequate for queries which are specific and empirical in nature. If, on the other hand, retrieval systems provide a link to represent the relationships between the variables of interest as reported in the documents, queries ... would be better answered. That is, precision might be enhanced for specific and empirical queries when the relationships between the index terms were specified in retrieval systems" (Oh, 1998, p. 290).

There is a growing body of work (e.g., Chen, 2002; Hetzler et al., 1998; Wong et al., 2000) 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 (e.g., Chen & Paul, 2001; Wise et al., 1995).

Appears in:

Proceedings of the American Society for Information Science & Technology Annual Meeting (pp. 52-58), 10/2003.
copyright ©2003 Information Today.

Visual representation provides an appropriate solution to the challenge of maintaining the interrelationships between documents and research concepts, can assist in understanding conceptual relationships across a domain, and can even assist in making discoveries (Benoit, 2002).

Putting it All Together

While other research has demonstrated that aggregating and analyzing research findings across domains augments knowledge discovery, 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 integrates three components to retrieve, display, and summarize research reports across a domain: 1) a Research Report Schema which contains research methods and findings extracted from domain documents, presented in a consistent, coherent and structured schema format that functions as a document surrogate to facilitate searching as well as rapid review of retrieved documents; 2) Research Concept and Relationship Extraction that incorporates controlled vocabulary based on the Unified Medical Language System (UMLS) to index the research findings extracted from data tables and figures; and 3) a Visual Exploration Interface which provides a dynamic map of research findings that graphically displays what is known as well as, through gaps in the map, what is yet to be discovered.

Telemakus is funded by the Ellison Medical Foundation and is a component of the Science of Aging (SAGE) project in partnership with the American Association for the Advancement of Science and Highwire Press at Stanford University to create an online resource for researchers in the field of aging (<http://sageke.sciencemag.org/>). The SAGE mission is threefold: deliver high-quality information on research in the field of aging and related disciplines; provide tools for more efficient searching and retrieval of information; and create a setting where researchers feel encouraged to share information and engage in discussion. The initial Telemakus database domain and visualization model represents knowledge related to caloric restriction and nutritional aspects of aging, a subset within the domain of the biology of aging—although Telemakus tools and visual mapping algorithms are broadly applicable to any domain that presents research findings as numeric data (i.e., in tables and/or figures). Caloric restriction was an ideal starting point for Telemakus because it is an important rapidly expanding specialized area of the biology of aging that is also highly interdisciplinary.

Fig. 1. Research Concept and Relationship Extraction.

Research findings (as concepts) and their relationships are derived from the tables and/or figures in the research report. Extracted research concepts are used to index the documents this allowing researchers to search the database for studies that report a statistically significant relationship. We are using the UMLS Metathesaurus (META) as the basis for creating a controlled concept vocabulary. The UMLS project is a large-scale National Library of Medicine (NLM) research effort to develop knowledge-based tools and resources to compensate for differences in the way in which concepts are expressed in the field of biomedicine. META is a database of information on concepts that appear in one or more of a number of different controlled vocabularies and classifications. It provides a uniform, integrated distribution format from over 95 biomedical vocabularies and classifications and contains syntactic information (NLM, 2003). 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). The current Telemakus system relies heavily on the UMLS for vocabulary control throughout the knowledgebase. An example of a source figure and its extracted concepts and relationships is shown in Fig. 1 above.

The schema includes standard bibliographic information (author, title, journal), information about the research design and methods (age, sex, number of subjects, pre-treatment and treatment regimen, organism and source of organism), and, most importantly, research findings derived from data tables and figures. Figure 2 is an example Research Report Schema from a query for documents reporting a concept relationship that includes "neoplasms" (i.e., the report must include a table or figure with numeric relationships of which neoplasms is one of

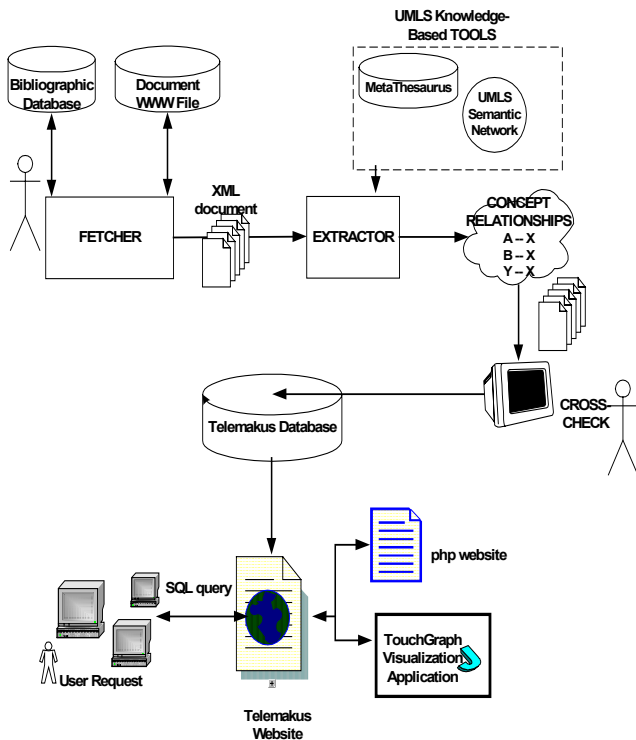


Fig. 4. Telemakus System Architecture.

Document Processing & Database Building

The purpose of the document processing and database creation components is to automate the addition of relevant parts of research articles to the database. 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.

The document processing system runs as server code behind an Apache server on a dedicated machine. It currently consists of the following three discrete phases: Fetcher, Extractor and CrossCheck, each independently responsible for its own task. We chose to use the Extensible Markup Language (XML) to cohesively pass data through the subsystem, thus minimizing any coupling of the individual components, as well as to provide formatted output to other parties. This enables the use of third-party packages to assist in collecting and processing the data, and easy reconfiguring via protocols like XMLRPC, to remote servers as the need arises.

Fetcher is responsible for gathering all the resources necessary to process a given research report that are needed by the subsequent phases. Fetcher takes two inputs: the document URL and its bibliographic database information

(citation). After creating an XML document and a unique directory on the server, Fetcher downloads the report 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 is responsible for extracting 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 takes as input the XML document from Extractor and 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.

Interactive Mapping Tool & User Interface

An open-source Java-based graphing applet, TouchGraph (<http://www.touchgraph.com>), has been adapted and enhanced for use as a component of the Telemakus user interface. The tool creates dynamic maps of research findings based on the research concept queried either via the schema or retrieval set. The tool was chosen for Telemakus because of its flexibility, customizing capabilities, high-quality source code, and compatibility with most browsers and operating systems. 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. Further, tool bars permit narrowing and broadening of the focus so that fewer or more research relationships are viewed; in addition, rotation of maps for improved viewing is supported.

Evaluation & Feedback

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. The primary aims of this testing were to:

Appears in:

Proceedings of the American Society for Information Science & Technology Annual Meeting (pp. 52-58), 10/2003.
copyright ©2003 Information Today.

1. Understand how knowledge domain specialists prefer to work with the research literature.
2. Model preferred features based on #1.
3. Test the completeness of schema elements and structure as a document surrogate.
4. Experiment with and identify optimal visual representations to meet user needs.
5. Iteratively review/evaluate/test for improved performance in response to user feedback.

In general, response to each successive version of Telemakus has been positive and included constructive feedback for system enhancements and expansions. User feedback affirms that retrieval based on research findings is a unique and highly desirable core function. Further, the Telemakus schematic document surrogate has been enthusiastically received as a major improvement over the traditional citation format with abstract. As one researcher stated, "The strengths of Telemakus are doing what PubMed does not do, which is to give an outline of the main points, and to allow searching over the figure/table legends, organisms/sources and outcome fields." Because basic sciences researchers tend to hone in on the data found within a report's tables and figures (sometimes before actually reading the article), extracting the headings and providing linked research concepts mimics a researcher's traditional approach to reading the research literature.

Next Steps

Our users have responded positively to Telemakus and 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, as well as personal needs (e.g. color blindness). Also, as more concepts are present the display becomes increasingly crowded and confusing (the "hairball" effect). Users need a variety of ways to prune the number of concepts so the presentation remains meaningful.

Speeding up document processing so Telemakus can easily and efficiently scale for comprehensive treatment of domains is the highest priority. The next steps are to improve and automate all components of the pipeline for adding documents and extracting schema elements and expand the concept-mapping component with cues to focus and visualize patterns in support of discovery of new knowledge. We are currently expanding use of the UMLS META resources in order to capitalize on UMLS Semantic Network for enhanced searching capabilities and are experimenting with MetaMap—a NLM program based on symbolic, natural language processing (NLP) and computational linguistic techniques to map biomedical text to the UMLS META—to parse the headings of tables and figures into UMLS terms (Aronson, Rindflesch & Browne, 1994). We are also experimenting with creating filters for viewing maps based on a variety of criteria, including

Semantic Type or organism studied (e.g., human versus rat) as a means of overcoming the "hairball" effect.

To accomplish these improvements, we plan to make several enhancements to existing services and significant changes to the mapping applet. Touchgraph will be extended to handle the concept of sets within a graph and the layout algorithms will be extended to group sets visually. The DTD for the XML document exchanged between the mapping servlet and Java applet will be extended to handle more complex networks, namely sets of nodes within the graph. Node entities will gain a new optional 'set' attribute defining which set they belong to. The database already contains all the relational links defining what semantic types are linked to each research term. The mapping servlet will need to populate the XML document with this additional data and the relationships between the semantic types (as defined by the UMLS Semantic Network) and the research terms.

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 the remarkably useful online bibliographic databases, the results of database searches 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 concept-based information system provides a flexible new method for exploration and knowledge discovery in a database of research reports. In 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 are grateful for the support of the Telemakus team, including Craig Benson, Heather Fuller, Wendy Kramer, Lucas Reber, Lisa Tisch and Jerome Woody. We also wish to thank our anonymous testers for their time and helpful feedback. Telemakus is funded by the Ellison Medical Foundation.

REFERENCES

- Aronson, A.R., Rindflesch, T.C. & Browne, A.C. (1994). Exploiting a large thesaurus for information retrieval. In Proceedings of the RIAO (Computer-Assisted Information Retrieval) Conference (pp. 197-216).
- Benoit, G. (2002). Data discretization for novel relationship discovery in information retrieval. *Journal of the American Society for Information Science & Technology*, 53, 736-746.
- Chen, C. (2002). Visualization of Knowledge Structures. In S.K. Chang (Ed.), *Handbook of Software Engineering and*

Appears in:

Proceedings of the American Society for Information Science & Technology Annual Meeting (pp. 52-58), 10/2003.
copyright ©2003 Information Today.

- Knowledge Engineering, Vol. II. Emerging Technologies (pp. 201-238). World Scientific Pub. Co.
- Friedman, C., Kra, P., Yu, H., Krauthammer, M. & Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*. 17, S74-S82.
- Fuller, S. (1983). Schema analysis: an approach to content representation of clinical trial reports. In *Proceedings of the 46th American Society for Information Science Annual Meeting*.
- Fuller, S., Revere, D., Bugni, P., Reber, L., Fuller, H. & Martin, G.M. (2002). Modeling a concept-based information system to promote scientific discovery: The Telemakus System. In *Proceedings of the American Medical Informatics Association Annual Symposium* (p. 1024).
- Hetzler, B., Harris, W.M., Havre, S. & Whitney, P. (1998). Visualizing the full spectrum of document relationships. In *Proceedings of the 5th International ISKO Conference: Structures and Relations in Knowledge Organization* (pp. 168-75).
- Khoo, C. & Myaeng, S.H. (2002). Identifying semantic relations in text for information retrieval and information extraction. In R. Green, C.A. Bean & S.H. Myaeng (Eds.), *The Semantics of Relationships: An Interdisciplinary Perspective* (pp. 161-180). Boston: Kluwer Academic Publishers.
- Kintsch, W. & van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review*. 85, 363-394.
- NLM. 2003 UMLS Knowledge Sources Documentation. 01/01/2003. Retrieved from: <http://umlsks5.nlm.nih.gov/kss/background/umlsReleases/2003AA/DOC/index.html>.
- Oh, S.G. (1998). Document representation and retrieval using empirical facts: evaluation of a pilot system. *Journal of the American Society for Information Science*. 49, 920-931.
- Piniewski-Bond, J.F., Buck, G.R., Horowitz, R.S., Schuster, J.H.R., Weed, D.L. & Weiner, J.M. (2001). Comparison of information processing technologies. *Journal of the American Medical Informatics Association*. 8, 174-184.
- Swanson, D.R. (2001). Information discovery from complementary literatures: categorizing viruses as potential weapons. *Journal of the American Society for Information Science*. 52, 797-812.
- Swanson, D.R. & Smalheiser, N.R. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*. 91, 183-203.
- Winograd, T. (1977). A framework for understanding discourse. In M.A. Just & P.A. Carpenter (Eds.), *Cognitive Processes in Comprehension* (pp. 63-88). Hillsdale: Erlbaum.
- Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A. & Crow, V. (1995). Visualizing the nonvisual: Spatial analysis and interaction with information from text documents. In *Proceedings of the IEEE Symposium on Information Visualization* (pp. 51-58).
- Wong, P.C., Foote, H., Leung, R., Adams, D. & Thomas, J. (2000). Data signatures and visualization of very large datasets. *IEEE Computer Graphics and Applications*. 20, 12-15.
- Wong, K.F., Song, D., Bruza, P. & Cheng, C.H. (2001). Application of aboutness to functional benchmarking in Information Retrieval. *ACM Transactions on Information Systems*. 19, 337-70.

Appears in:

Proceedings of the American Society for Information Science & Technology Annual Meeting (pp. 52-58), 10/2003.
copyright ©2003 Information Today.