M&M-Query: Database support for the annotation and retrieval of biological research articles

Kenneth Baclawski and Natalya Fridman*
Northeastern University
College of Computer Science
Boston, Massachusetts 02115
(617) 373-4631
FAX: (617) 373-5121
{kenb, natasha}@ccs.neu.edu

December 6, 1993

Abstract

The Biological sciences produce an enormous research literature every year. Research papers are highly structured documents which are not captured using the traditional techniques of information retrieval: keywords and flat text. This is especially true of the Materials & Methods section of experimental papers. This report describes a prototype system for using database and text analysis techniques for making this literature more easily accessible. The system integrates a biological knowledge base with structured documents in an object-oriented database. The system has a graphic user interface that supports a variety of techniques for expressing queries. This prototype is the "proof of concept" for a more elaborate system called the Ontology Builder which will provide support for constructing and maintaining subject-specific ontologies.

1 Introduction

Biology is a very large and diverse field. The primary output of the enterprise is its published research literature, which consists of about 600,000 papers every year. The vast majority of these papers report experimental work and do so in a highly structured manner[BFH+93]. This report and video presents a prototype system for using database and text analysis techniques for making this literature more accessible.

*This material is based upon work supported by the National Science Foundation under Grant No. IRI-9117030.

As the name of the prototype suggests, the system emphasizes the Materials & Methods sections of biological research papers. There are a number of reasons why we chose this particular section.

- The section is an important one. Biology is a technique-driven rather than a theory-driven experimental science. Nearly every biological research paper has a Materials & Methods section.
- This section is easier to analyze than the rest of the paper. It has a more limited (although still large) vocabulary. The language used is more stylized (although not simple) and is primarily descriptive with no arguments, discussion of hypotheses, etc. The other parts of the paper use more complex syntactic and linguistic forms.

The system integrates a biological knowledge base with structured documents in an object-oriented database. The system has a graphic user interface that supports three methods of expressing queries: natural language, keyword search and knowledge frame fill-in. The frame fill-in interface is also used for annotating the documents in the database using knowledge frames.

The M&M-Query System is a prototype for a more ambitious tool that we call the Ontology Builder. Eventually many features of M&M-Query that are now "hard coded" will be generated automatically from a higher-level specification that we call an "ontology." The word ontology literally means "a branch of metaphysics relating to the nature and relations of being." Our use of the word is much more restrictive, dealing only with the nature of, and relationships among, concepts within a narrow subject area. Attempts to specify ontologies for scientific disciplines are very common, with most disciplines having some kind of subject classification scheme by this time. However, as Lakoff points out[Lak87], "human categorization is based on principles that extend far beyond those envisioned in the classical theory." As a result, simple classification methods leading to taxonomies of concepts are inadequate for expressing the rich variety of human categorization techniques.

The Ontology Builder will allow groups of individuals working in a single well-defined discipline to construct a subject-specific ontology for their field that is significantly richer than simple taxonomic classification. Having constructed their ontology, the Ontology Builder will then automatically generate tools for using the ontology. The Ontology Builder will integrate techniques and methodologies from such disparate areas as Object-Oriented Database Management, Artificial Intelligence and Information Retrieval.

2 Related Work

Recent technical advances in molecular biology have allowed the generation of an enormous volume of data, which cannot be dealt with by traditional printed publications. A second form of electronic data publishing has been developed to make data available in a world-wide network-accessible database, while methods and conclusions continue to be published in traditional publications [CFGB91]. Databases such as GenBank [BCF$^+$92] and The EMBL

data library [HFSC92] serve as repositories for DNA sequence data; PIR [BGHG91] and SWISS-PROT [BB91]store protein sequence data; and GDB stores human gene mapping information [PMFR92]. Data in these databases are now accessible for computer analysis. The content of research publications, which describes the methods and conclusions associated with the data, remains inaccessible to sophisticated computer search. Our efforts are directed toward making the contents of publications accessible by applying database techniques to the actual content of research publication.

The results of [LS91] support the claim that systems incorporating natural language processing techniques are more effective than systems based on stochastic techniques alone. An example of applying natural language processing techniques in a limited domain to extract the information from flat text is in [JR90]. This system used a combination of bottom-up and top-down parsing to analyze news articles and to retrieve information about financial companies.

The EDS TemplateFiller system [SMHC93] applies Message Understanding (MUC) text-filtering techniques to the generation of knowledge frames for one or a few specific subject areas from entire texts (computer product announcements). TemplateFiller fills in slots for frames that exist in a predefined schema of templates, ignoring subjects that are not in the schema.

The Material & Methods section is similar in some ways to a very complex recipe. The Recipe Acquisition System developed in the University of Connecticut [MMP92] attempts to understand the recipes in a particular Chinese cookbook. Their work differs in a number of ways from ours. They try to fully understand the recipe but make no provision for queries. In our project, the emphasis is on queries. Their project is not designed to evolve and is unlikely to scale up very well. Our project must both evolve and scale up.

The Unified Medical Language System (UMLS) of the National Library of Medicine (NLM) [LHM93, HL93] is an example of a subject-specific ontology that has gone significantly beyond simple taxonomic classification. The UMLS is developing a Metathesaurus and Semantic Network to support consistent retrieval of electronic biomedical information from a variety of sources such as bibliographic or factual databases or expert systems. The UMLS compensates for variations in the way similar concepts are expressed in different sources, and for the scattering of useful information among disparate computer systems. The scope of the UMLS is broader than just information retrieval, but its approach is closely related to ours.

3 System Description

3.1 General Description

The M&M-Query System is designed to assist biologists in finding articles in the research literature. The first and largest component of the database is the corpus of research papers that are stored as marked up text using the Standard Generalized Markup Language (SGML)[A+86]. Storing the text in marked up form permits the system to display the pa-

per as it would be seen in published form. This is especially important in biology where typesetting details (such as font changes) convey semantic information.

The other main component of the database is the biological knowledge base. This consists of a schema, a lexicon and annotations of papers in the corpus[BFFP93]. The schema defines the structure of knowledge frames that are used to annotate[1] each paper and to express the system's understanding of a query. These annotations are mainly based on the information in the Materials & Methods section. This section is essentially a large and complex recipe describing how various input materials are transformed ultimately into output measurements. As a result there are two kinds of frames: materials and processing steps. Materials are either initial ingredients or intermediate materials, which are classified into categories such as protein, plasmid and so on. Processing steps are more complex since they consists of a connected sequence of steps that transform input substances into output substances. Furthermore, a process step can be elaborated into subsidiary steps, which in turn can be elaborated down to several levels. At the lowest level, an elementary step is defined using a small set of parameters such as the temperature, duration, etc. In addition to specifying a material or process step in terms of parameters or other materials and steps, one can use a standard name which identifies one of the objects in the lexicon.

3.2 User Interface

The principal use of the system by a biologist would be for information retrieval: to find those papers that deal with particular materials, processing steps, or a combination of the two. The user interface allows for several kinds of query: natural language queries, choosing topics from a traditional concept classification, or by filling in knowledge frames.

The result of a query is a list of papers that are relevant to it. The user may then view the text of any particular paper in the list. A useful feature of any good information retrieval system is some mechanism for explaining why each paper was included in the result of a query. The M&M-Query System does this by highlighting the parts of the paper which were responsible for the paper being retrieved. In addition to explaining why the paper was retrieved, it also focuses the user's attention on the portions of the paper that are likely to be the most interesting.

Perhaps the most interesting way of stating a query is the frame-filling mechanism. This mechanism is the one most closely related to the underlying ontology. Whatever mechanism the user uses to formulate a query, it is converted into knowledge frames that can be directly manipulated using frame-filling. Thus a user can formulate their query initially using natural language, and then, if the system doesn't seem to be understanding the query, the user switches to the frame representation to see if some term was misinterpreted.

Frames and slots are labeled with standard biological terms and for each slot one can look at a list of values that occur in the corpus of documents. This can help users formulate queries quickly and precisely. Slots need not just be values, they can also be another complex

[1]We are misusing the word "annotate" slightly here. Although our annotations can refer to portions of the document, they classify rather than explain these portions of the document.

object. For example, a material that is used as part of a process step might itself have to be prepared using a series of steps. The graphic user interface deals with this situation by opening another window that partially obscures the first window. The elaboration on the slot value can then be done within this new window, which itself can have another detail window, and so on.

The ontology was created via consultations with molecular biologists so that it reflects their view of the field. In the later versions of the system we are planning to allow the users to modify the existing ontology schema and, therefore, completely customize it to their own view of their discipline.

3.3 Annotating Papers

The M&M-Query System may be used not only for making queries but also for entering information about a paper that is then used for retrieving it. Each paper has a collection of knowledge frames associated with it called a semantic keynetwork (or simply *keynet*)[BS93]. A keynet is analogous to the set of keywords from a subject-specific concept classification that is currently used in most research journals. The keynet is more than just a collection of subject classifiers for each paper as a whole. It is semantically richer, and keynet frames and slots can refer to specific parts of the document.

The keynet for a paper could be generated by the author of the document, assuming good tools were available. Such a task would be no more effort than is now required for writing the abstract or selecting the keywords. While it may eventually be possible to use natural language processing techniques to generate keynets, this is only possible for textual information objects. A third possibility is to have professional annotators construct the keynets. This is less costly than one might expect. Based on our experience with the construction of keynets by professional biologists, it would take less than an hour to write a high-quality keynet for a biological research paper. This works out to less than $30 million per year to annotate the entire biological research literature, a tiny amount compared to the cost of generating this literature in the first place.

The same kinds of frames are used both for queries and for keynets, except that query frames do not have the ability to refer to any other text. When a document is retrieved, it is the result of one or more matches between parts of the query and parts of the paper's keynet. The references to the parts of the document allow the M&M-Query System to highlight the parts of the papers that were responsible for the paper being retrieved.

4 Future Work

As mentioned earlier, the M&M-Query System is a prototype for the Ontology Builder. Its purpose is to support the development of subject-specific ontologies that are semantically richer than simple taxonomic classification systems. To achieve this, the main components of an ontology are:

- Schema. This defines the underlying structure of knowledge frames for this subject. It consists of class definitions, each of which consists of a set of attribute definitions. Typical subject-specific schemata have around 200 classes.

- Lexicon. This component contains lists of well-known objects that are instances of classes in the schema. Lexicons can be very large, containing on the order of 100,000 terms.

- Thesaurus. Semantic relationships among classes and objects are defined in this component, which consists of a schema-level subcomponent and a lexicon-level subcomponent. Each relationship has an associated behavior that controls, for example, whether a class or object can replace another ("broadening") and how such a replacement affects the attributes of the class or object. Note that semantic relationships are not the same as structural relationships. Structural relationships are used as a convenience for organizing classes that have attributes in common but that could be unrelated semantically. Semantic relationships specify links of various strengths between classes and objects which need not be reflected by any common attributes at all.

- Sublanguage. Classes and objects have associated natural language representations. On output, the representation consists of "stock phrases" for expressing the classes and objects. Input is the more complex operation because of the large variety of syntactic forms that can be used to express a concept.

An example of such a semantically rich system is the UMLS that was described in the Related Work section above. However, the UMLS has no GUI for entering knowledge frames, there is no mechanism for modifying the UMLS ontology, there is no integration with a database management system for storing knowledge frames, and the UMLS has no behavior associated with any of its classes or relationships.

The central module of the Ontology Builder is the Design Tool used for designing the schema. We are developing new graphical methods for schema design which could be used by someone without a background in Computer Science, but still give the means to represent the complex concepts which exist in the real world. An important part of this complexity is behavior. Most graphic schema design tools deal primarily with structure and, except for inheritance, do not represent behavior. The Schema Design Tool will use techniques from GUI design to allow users to specify other forms of substitutions and inference than just inheritance, such as synonym, part of, defaults and so on. Representing such behavior graphically using familiar, easy-to-understand graphical forms is the research challenge we are currently facing.

Having designed an ontology, the Ontology Builder will automatically construct a variety of representations and tools:

1. Traditional graphical representations. Some users have already seen the traditional schema representations using boxes and arrows, and for such users this can be a useful conceptual device.

2. Natural language representation. Those who have no experience with graphical representations of schemata can simply read a natural language description of the schema which is generated using stock phrases.

3. The Frame Fill-In Tool. This tool is discussed in subsection 3.2 above.

4. The Object-Oriented Database. All the information, of course, has to be stored. And the best way to store all the objects, attributes and relationships is in an object-oriented database. The database schema will therefore be an important output of the Ontology Builder.

5. Semantic Network Translator. This module translates between various knowledge representations. For example, a slightly different representation will be needed by by high-performance search engines such as the `KEYNET` information retrieval system[BS93].

Acknowledgements

We would like to acknowledge all of the faculty and students who participated in the design and development of the M&M-Query prototype: Robert Futrelle, Carole Hafner, Maurice J. Pescitelli and Chendong Zou.

References

[A+86] Anonymous et al. Information processing – text and office systems – standard generalized markup language (SGML), 1986. ISO 8879-1986 (E).

[BB91] A. Bairoch and B. Boeckmann. The SWISS-PROT protein sequence data bank. *Nucleic Acids research*, 19, Supplement:2247–2249, 1991.

[BCF+92] C. Burks, M.J. Cinkosky, W.M. Fischer, P. Gilna, J.E. Hayden, G.M. Keen, M. Kelly, D. Kristofferson, and J. Lawrence. GenBank. *Nucleic Acids research*, 20, Supplement:2065–2069, 1992.

[BFFP93] K. Baclawski, R. Futrelle, N. Fridman, and M. Pescitelli. Database techniques for biological materials & methods. In *First Intern. Conf. Intell. Sys. Molecular Biology*, 1993. to appear.

[BFH+93] K. Baclawski, R. Futrelle, C. Hafner, M. Pescitelli, N. Fridman, B. Li, and C. Zou. Data/knowledge bases for biological papers and techniques. In *Proc. Sympos. Adv. Data Management for the Scientist and Engineer*, 1993. to appear.

[BGHG91] W.C. Barker, D.G. George, L.T. Hunt, and J.S. Garavelli. The PIR protein sequence database. *Nucleic Acids research*, 19, Supplement:2231–2236, 1991.

[BS93] K. Baclawski and J. E. Smith. KEYNET: Fast indexing for semantically rich information retrieval. In *Proc. ACM SIGMOD Conference*, 1993. Submitted.

[CFGB91] M.J. Cinkosky, J.W. Fickett, P. Gilna, and C. Burks. Electronic data publishing and GenBank. *Science*, pages 1273–1277, May 1991.

[HFSC92] D.G. Higgins, R. Fuchs, P.J. Stoehr, and G.N. Cameron. The EMBL data library. *Nucleic Acids research*, 20, Supplement:2071–2074, 1992.

[HL93] Betsy L. Humphreys and Donald A.B. Lindberg. The UMLS project: making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association*, 81(2):170, Apr 1 1993.

[JR90] P. Jacobs and L. Rau. SCISOR: Extracting information from on-line news. *Comm. ACM*, 33:88–97, November 1990.

[Lak87] George Lakoff. *Women, Fire and Dangerous Things*. The University of Chicago Press, Chicago, IL, 1987.

[LHM93] D.A.B. Lindberg, B.L. Humphreys, and A.T. McCray. The Unified Medical Language System. *Methods of information in medicine*, 32(4):281, Aug 1 1993.

[LS91] W. Lehnert and B. Sundheim. A perfomance evaluation of text-analysis technologies. *AI Magazine*, 12(3):81–94, Fall 1991.

[MMP92] R. McCartney, B. Moreland, and M. Pukinskis. Case acquisition from plain text: reading recipes from a cookbook. Technical Report TR-CSE-92-20, University of Connecticut Department of Computer Science and Engineering, 1992.

[PMFR92] P.L. Pearson, N.W. Matheson, D.C. Flescher, and R.J. Robbins. The GDB human genome data base Anno 1992. *Nucleic Acids research*, 20, Supplement:2201–2206, 1992.

[SMHC93] H. Kelly Shuldberg, Melissa Macpherson, Pete Humphrey, and Jamil Corley. Distilling information from text: The EDS TemplateFiller system. *Journal of the American Society for Information Science*, 44(9):493–507, 1993.