User-Centered Biological Information Location by Combining User Profiles and Domain Knowledge

Xindong Wu [(+)], Jeffrey E. Stone [(+)], and Marc Greenblatt [(*)]
[(+)] Department of Computer Science [(*)] Department of Medicine
University of Vermont, Burlington, VT 05405

Abstract

To aid researchers in obtaining, organizing and managing biological data, we have designed an intelligent digital library system that utilizes advanced data mining techniques. Our digital library system is implemented as a centralized J2EE web application with links to publicly accessible data repositories on the Internet. The digital library is based on a framework used for conventional libraries and an object-oriented paradigm, and provides personalized user-centered services based on the user's areas of interests and preferences. To make personalized service possible, a "user profile" that represents the preferences of an individual user is constructed based upon a user's past activities, goals indicated by the user, and options. Utilizing these user profiles, our system makes relevant information available to the user in an appropriate form, amount, and level of detail with minimal user effort. The core of our project is an agent architecture that provides advanced services by combining data mining capabilities with domain knowledge in the form of a semantic network. The semantic network imparts a knowledge structure through which the system can "reason" and draw conclusions about biological data objects and provides a federated view of the many disparate databases of interest to biologists. In the development of our semantic network, we have included the concepts from several established controlled vocabularies, chief among them being the National Library of Medicine's Unified Medical language System (UMLS). Our complete semantic network consists of 183 semantic types and 69 relationships

1 Knowledge Object Modeling

Biological information can be broken down and represented in forms of objects with the integration of logic rules. Each object can have a set of rules that govern its behavior and appearance, as well as communication to other objects. Association links between objects can be represented in the form of rules in knowledge objects and used to conduct heuristic search over the databases.

With knowledge object modeling [Wu & Cai 2000], a biological item is decomposed into knowledge objects (classes or objects for short), each holding an internal state and a well-defined behavior. The object's internal state is defined by attributes and constraints, and its behavior by rules and methods. The biological item object represents a variety of biological data including annotation and publications, and is constructed by the basic building block of description. Instances of the biological item class at the top level of the hierarchy will correspond to the most general form of biological data. This hierarchy will be designed to completely cover all types of biological data in digital libraries, including images and text, and ranging from page images to interactive and compound documents. The biological item object can be composed of some description objects and some media objects, or just one or the other.

A significant problem for most biological databases and libraries is that of annotation. The detailed information needed to describe biochemical processes and genetics is much more complex than the information of the taxonomy of living species that Linnaeus developed in the 18th century. We need ways to transfer this information efficiently and without propagating error. Our design is based on the biological structures embodied in the target databases and is represented as a class hierarchy. At the top is the most general knowledge concept that could be decomposed into detailed and related knowledge modules. Each node in this structure is represented as a biological item introduced above.

Biologists are interested in specific genes and their products. Therefore a biological object is divided into genetic information, gene ontology, and structure information. Each of these classes is further divided into more detailed classes that give the finer details of biological information. The genetic information will contain the DNA sequence, the gene's location in the genome, and its homologues. The Gene ontology information comes from the Gene Ontology

Consortium and contains information about molecular functions, biological processes and cellular components. The structure information will include raw data, electron density, and structure annotation. Other information provided will be text, references, history and information about possible relations to diseases. With this approach we attempt to outline a biologist's view of the gene.

The design of the knowledge structure both allows the depiction of the overall knowledge underlying the biological process and facilitates the system to efficiently search the right information in the databases. This structure helps a more focused, context-based retrieval over the dispersed databases.

2 The Semantic Network

In the construction of a dictionary to specify the scope of our digital library, we were presented with several difficulties due to the complex nature of biological data. These problems include multiple names for the same entity, the dependency of the biological state in which the function is taking place, and multiple functions for the same protein. To overcome these difficulties and to add additional functionality to our digital library, we have developed our dictionary as a semantic network.

Although there have been several ontologies developed for describing biological data, there is still no published knowledge base that can be used to cover the number of disparate databases which are used by biomedical professionals. Yu et al (1999) adapted the UMLS semantic network to cover genomic knowledge and Hafner et al (1994) also used the UMLS as a basic building block for their system of representing biomedical literature. Most other biomedical resource systems such as Genbank and the Protein Data Bank (PDB) contain crucial facts, but do not contain information about the concepts and relationships of the many inter-related terms (PDB).

The Gene Ontology Consortium has developed a large controlled vocabulary for the unification of genetic concepts and terminology. This controlled vocabulary along with several others is now part of the massive UMLS Metathesaurus. These ontologies provide the vocabulary for the description of many biological concepts such as the annotation of the molecular function, biological process, and cellular component of gene products. This metathesaurus is a big step towards the unification of biological knowledge, however, it is simply far too complex to provide a federated solution to unifying biological databases.

The structure of the Gene Ontology vocabulary provides a good example of the vocabularies that make up the UMLS Metathesaurus. The Gene Ontology controlled vocabulary is based on the annotation of gene products. A gene product is a physical entity. Gene products may be RNA or proteins. These gene products may have many molecular functions. A molecular function is a description of what a gene product does. One drawback of the Gene Ontology system is that the molecular function only describes what a gene product has the potential to do without regard to where or when this function may take place. Such semantics as to where and when a function takes place could be contained within a semantic network.

The National Library of Medicine has a long term project to build a Unified Medical Language System (UMLS) which is comprised of three major parts: the UMLS Metathesaurus, SPECIALIST Lexicon, and the UMLS Semantic Network. The Metathesaurus provides a large integrated distribution of over 100 biomedical vocabularies and classifications. The Lexicon contains syntactic information for many terms, component words and English words, including verbs, not contained in the Metathesaurus. The Semantic Network contains information about the types or categories to which all Metathesaurus concepts have been assigned and the permissible relationships among these types (UMLS). The UMLS system has been used successfully in many applications mostly involving scientific literature.

The UMLS Semantic Network provides an ideal framework for federating disparate databases. However, the current structure of the UMLS Semantic Network is most useful for scientific literature and clinical trial information. If one is trying to use the UMLS Semantic Network for federation of several disparate databases, they will find the network is not sufficiently broad to cover the multiple items in all of these databases.

We have therefore decided that to best suit the needs of our digital library system, we must develop our own controlled language system. To do this, we started with the basic framework of the UMLS semantic network and then pruned some of the less important details and added new concepts and relationships where needed to cover the databases in our digital library.

3 User Profiling and Recommendation Agents

Capturing user preferences can be a difficult task. Simply asking the users what they want can be obtrusive and error prone. In fact, the user might not even know what they really want. On the other hand, monitoring user behavior may be unobtrusive but can also be computationally time consuming and discovering meaningful patterns is difficult. Yet capturing user information is critical for the recommender system of our digital library.

User profiling methods can be classified as either knowledge based or behavior based. Knowledge based methods employ questionnaires or interviews to dynamically match users to one of a number of different static models of users.

Behavior based approaches seek to capture the user's behavior and apply data mining techniques to discover useful patterns in the behavior. This approach will need to log the user's behavior in some manner.

The user profiling employed by recommendation agents is primarily behavior based. With most recommendation agents a binary, two-class model that represents what a user likes or dislikes is used. Data mining techniques are then used to discover meaningful information about the user. In our system this meaningful information is in the form of rules. The recommendation agent will then use these rules to recommend items that she may be interested in.

The user knowledge can be collected either implicitly or explicitly. Implicit knowledge acquisition would be the preferred mechanism since it has little or no impact on the user's normal work. Analyzing the click stream as a user navigates through our system might provide one method for collecting this information in an unobtrusive manner. This type of knowledge acquisition requires some degree of interpretation to understand the user's real interests. This is an inherently error prone process. How do we, for instance, determine if a user is lingering on one item because they are truly interested or if they were interrupted while navigating the site?

Explicit knowledge acquisition, on the other hand, requires the user to interrupt their normal work to provide feedback. This may be undesirable, but will generally provide the system with high confidence information since the user themselves provides the information. This feedback is most often in the form of a questionnaire on the relevance, interest and quality of an item. It may also come in the form of programming where the user is asked to create filter rules either visually or via a programming language.

Our system utilizes a combination of these different systems. When the user performs a search on our system, they may mark items as interesting. These items are then saved in the user profile. At any time the user may choose to use this profile to generate new rules. Due to the imprecision of this method, there is a third step where the user provides additional feedback on which rules to add to the profile.

4 System Architecture

There are many factors involved in determining the architecture of a digital library. In making this decision, one must determine how the system will be used and who the users will be. We want the system to be robust and scalable, but we also have to face the reality of a limited budget. We also want the system to be available to users throughout the university campus and also users from other institutions. The system must also combine several different agents together in a seamless manner, some of which may be difficult to modify.

Based on an analysis of the use of a system such as ours, we decided that a web application would be the ideal architecture for our digital library. Since the Weka data mining package was written in Java, we decided on a J2EE application server. We are using the Tomcat server to host our web application. This server is a free, open-source system available under the GNU Public License and is the reference implementation of the servlet 2.3 and JSP 1.2 specifications. Tomcat is quite powerful as either a stand-alone web server or embedded within an Apache server.

Tomcat can recognize standard HTML files, Java Server Pages (JSP) or Java servlets. Servlets are Java technology's answer to Common Gateway Interface (CGI) programming. They are programs that run on a Web server, acting as a middle layer between a request coming from a Web browser or other HTTP client and databases or applications on the HTTP server. It can be argued that Java servlets are more efficient, easier to use, more powerful, more portable, safer and cheaper than traditional CGI and other technologies. Java Server Pages allow us to include Java code inside of an HTML page. This provides the author with close control over the web design.

The primary feature that sets our system apart from other systems is the recommendation agent. This agent will generate rules about the users and learn about the users' interests and preferences. The rules will be refined and improved through two learning processes: interactive incremental learning and silent incremental learning. Our system will first learn about a user's areas of interest by analyzing the user's declared interest topics and the user's visit records, and then assist the user in retrieving the right information. The user profile is composed of a set of biological terminologies coming from the knowledge structure and the dictionary. Interactive incremental learning will function in cycles that interact with the user. The system will prompt the user with a set of related documents which are likely of the user's interest, and ask for feedback on the level of interest in each of these documents. Considering the feedback, the system will make changes to its search and selection heuristics and improve its performance.

Our system works differently from search engines and other kinds of agents like WebWatcher [Joachims et al 1995] and [World Wide Web Worm] that help the user on the global Web. First, through incremental learning of the user's characteristics or interest areas, the system will become an assistant to the user in retrieving relevant information. Second, our library will have the potential to reduce user accessing and retrieval time, by displaying a list of changes that have been made since the user's last visit. Finally, the system can be easily adopted for other digital libraries. This can be accomplished by adding a different knowledge source for the dictionary.

5 Future Research

The portion of the existing library that needs the most immediate attention is the database of items. This database is small at this time. We hope that this system will grow in the very near future to provide a valuable asset not only to the university research community, but indeed the world. Some of the research topics where future versions of this system may evolve are described below.

5.1 Sophisticated data entry for populating the library

At this time, the administrative page for logging items into the library is an html-form page. This process is very tedious and error prone. We would like to create an agent for the automatic entry of these items. Potentially some of the emerging web services will provide some tools or methods for entering this information directly from trusted sites.

5.2 Improvement in Semantic Searching

The current method for searching the semantic network for relationships is quite slow. This is due in part to the large size of the semantic network database (roughly 3 million entries). Some effort will need to be spent on improving the performance of this semantic search. Perhaps there may be better ways of implementing the semantic network via Java object creation in memory.

5.3 Emerging Semantic Web Technologies

One of the more exciting areas of research is in the semantic web technologies. New means of annotating web resources are promising to revolutionize the way we use the Internet. The current internet is designed mostly in markup languages to format this information for human consumption. With a new focus on web services and XML technologies, researchers are looking into ways to create a web designed by machines for machines.

Semantic Web technologies such as DAML and OIL could potentially be used for the generation of our semantic net [Berners-Lee 2001]. This would allow this net to extend beyond our own system to include other semantic networks and web services by searching the descriptors such as Resource Description Frameworks (RDF) for these systems for common nodes in the networks.

5.4 Expanding the Semantic Network

The current semantic network with knowledge objects contains information primarily designed for describing molecular biologists. Expanding this semantic network can be achieved by either adding new nodes to the semantic network, or by adding whole new semantic networks to the system. The modular design of our system allows one to plug in any semantic network that fits the general schema outlined in Section 3. This would allow our library to be used outside of its intended biological domain to any domain that one would want.

5.5 Additional User Agents

There are several additional user agents that could be developed for our system. Chief among these would be an update agent. This agent would notify the user of updates to items that fit their profile. The semantic network contains a valuable knowledge base that could be further exploited. An agent for navigating the semantic neighborhood of an item in a graphical map might also prove interesting.

References

[Hafner et al 1994] C.D. Hafner, K. Baclawski, R.P. Futrelle, N. Fridman, S. Sampath, Creating a knowledge base of biological research papers. In 2nd International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Stanford CA. 1994.

[Joachims et al 1995] T. Joachims, T. Mitchell, D. Freitag, and R. Armstrong, WebWatcher: Machine Learning and Hypertext, GI Fachgruppentreffen Maschinelles Lernen, K. Morik and J. Herrmann (Eds.), University of Dortmund, Germany, August 1995.

[World Wide Web Worm] http://www.cs.colorado.edu/home/mcbryan/WWWW.html

[Wu & Cai 2000] X. Wu and K. Cai, Knowledge Object Modeling, IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans, 30(2000), 2: 96-107.

[Yu et al 1999] H. Yu, C. Friedman, A. Rzhetsky, and O. Kra, Representing Genomic Knowledge in the UMLS Semantic Network, AMIA Symposium 1999, 181-185.