Getting Biologists to (Willingly) Do the Work of a Thousand Annotators

Shannon Bradshaw* **and Marc Light†**
*Department of Management Sciences
†Linguistics Department & School of Library and Information Science
*†Department of Computer Science
The University of Iowa
Iowa City, IA 52241
{shannon-bradshaw, marc-light}@uiowa.edu

Abstract

Biologists read large quantities of scientific literature. Their knowledge synthesis process involves extracting tables of facts, important passages, lists of relevant elements, etc. We describe the Machete system, now in development, that can both help the biologist extract such information and capture what has been extracted. In the process, Machete provides a means of curating scientific knowledge for a community of researchers with little or no addition effort.

Introduction

A number of large knowledge curation projects exist in the biological sciences. Two prominent examples are SwissProt and Mouse Genome Informatics (MGI). These curation projects focus on problems such as identifying the functions of genes and proteins, extracting semantic relations (e.g., protein interaction (Craven & Kumlien 1999)), summarizing the basic biology of a gene (Hersh & Bhupatiraju), etc. These curation efforts involve a combination of manual reading and markup and automated text-mining systems to help direct curators to passages of particular interest.

It is our hypothesis that provided with the right set of tools, biologists will perform many of these same curation efforts in the natural course of their information gathering activity. In biology, even with the availability of knowledge repositories such as SwissProt, the literature itself remains a primary source for the current state of human knowledge. One explanation for this may be that knowledge repositories are slightly behind the newest literature or that they don't contextualize the information they contain sufficiently to provide information seekers with complete answers to their questions. While we do not here posit a reason for this behavior, we hypothesize that biologists use the literature itself in some way to satisfy the majority of their information needs. In a preliminary study of information needs in the bioscience community, we together with several other researchers under the direction of William Hersch surveyed 43 bioscience researchers at 22 organizations to collect descriptions of their information needs and information gathering behavior (Hersh & Bhupatiraju 2004). The results of this study indicate that use of Entrez PubMed and Google to find relevant articles accounts for the majority of time biologists spend looking for information. It may be that information resources for scientists (including curated databases) will increase in utility if they are integrated with literature and basic literature search. This is the view we take in our work and to this end we are developing an information system called Machete (http://dollar.biz.uiowa.edu/∼sbradsha/Machete/). Machete is designed to capture, annote, and provide access to human knowledge entirely in the context of literature search tasks. With Machete, biologists extract, assemble, and document what is known about genes, proteins and other questions of interest in the normal course of their day-to-day literature review with little or no extra work. Using Machete, information seekers are directed to relevant information previously "curated" by other researchers with similar information needs.

The Machete System

In a recent genomic analysis of the poorly studied yet economically important red tide dinoflagellate species, *Alexandrium tamarense*, a graduate student found a gene encoding histone H2A.X. He suspected that this was an important finding but was unaware of the specific function of H2A.X or what had previously been reported about histones in dinoflagellates. A time consuming web search revealed that his finding was indeed newsworthy and worthy of in-depth analysis. H2A.X had never been reported in dinoflagellates and was in fact the first (bioinformatically) confirmed histone gene of any kind in these taxa. Further digging revealed that H2A.X is involved in DNA double-strand break repair in other model eukaryotes and could therefore play the same role in *Alexandrium*. In Machete, the integration of search, text-mining and information organization tools will enable researchers to more quickly identify avenues of promising research. Machete is built around the concept of a "knowledge artifact" (KA). As is depicted in Figure 1 A, a researcher will use the Machete client to search the Web as usual. For any documents viewed, users may ask that Machete automatically highlight relevant passages based on searches performed in a search session, find specific types of information such as which experimental methods were used, and ask for structured data to be extracted (e.g. a ta-

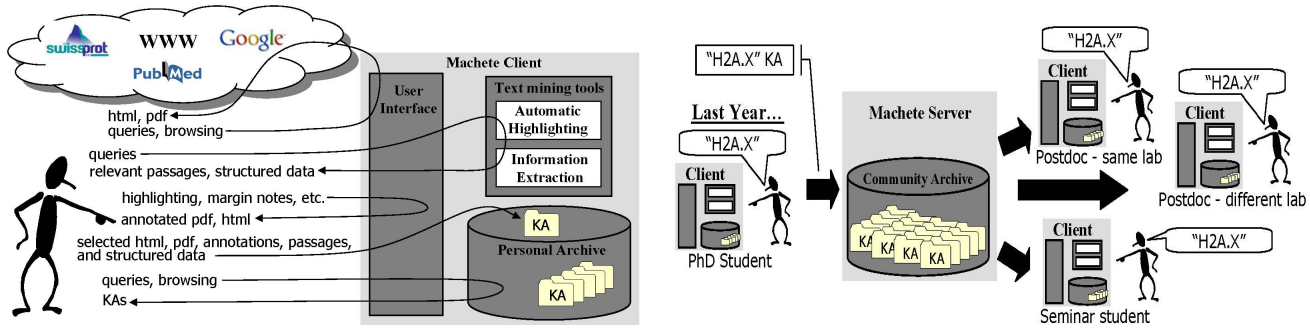

Figure 1: (A) Personal information management application (B) Knowledge management system

ble of protein interactions). A researcher may also highlight passages, add margin notes, and otherwise annotate on top of HTML and PDF documents. Finally, the researcher may save documents, highlighted passages, extracted information, etc. of his choice in an organized fashion in a KA. A variety of lenses through which users may view the assembled information will enhance a researcher's ability to analyze existing knowledge on a topic of interest. Returning to our example, Machete will allow researchers to generate KAs for all genes of interest in a non-model organism so that these can be rapidly screened for function or distribution to identify those of the highest interest (e.g., those involved in DNA packaging or repair as in *Alexandrium* H2A.X or in specific metabolic pathways).

This work need never be duplicated. Any KA generated for a gene in this example, may be reused by the same or other researchers in other work on that gene. In an information gathering task such as this it is unlikely the graduate student will remember all the genes for which he constructed KAs. Even if he did, another researcher in his lab probably would not know this and would likely not think to ask. Machete solves this problem by automatically distributing the KAs created by individuals to other members of a closed community (see Figure 1 B). Using a client/server approach, as a researcher begins an information gathering task, his client will begin polling the server for useful KAs created by other members of his community. When discovered, the client will inform the user in an unobtrusive just-in-time approach[1] (Budzik & Hammond 2000; Rhodes 2000).

Finally, KAs provide a means by which a group of researchers may collaboratively review the literature on a specific gene and make this information available to the wider bioscience community. Within a laboratory or other closed community, KAs will serve as vehicles for information and knowledge sharing among collaborators. Returning to the histone H2A.X example, a KA will provide a place to organize and save everything that a research group learns about histone H2A.X. Over time a KA will become an extensive review of the literature concerning a gene family or other topic of interest. Such KAs will be invaluable resources for other researchers with related information needs in the same way as a review article on the subject. Machete will enable researchers to edit annotations and add or remove highlighted passages, figures, documents, or structured information. In the bioscience community of today where review articles are so important, Machete provides a framework for generating information that is in effect much like a review article, but do so in the natural course of literature search with a fraction of the effort.

Training Data: A Natural Byproduct

A common approach to text processing tasks in bioinformatics is to build a system that can learn from examples that have been annotated manually. Many of these approaches make use of statistical machinery (but not all, see (Tanabe & Wilbur 2002) for a transformation-based learning approach). More broadly, supervised learning methods are a standard method in many areas of AI and are likely to play a role in future AI research. These methods often require large amounts of supervised data to achieve satisfactory performance. In almost all cases, acquiring such supervised data is difficult and expensive. This is particularly true in the biosciences, since annotators need graduate degrees to be able to do the annotation. See (Kim *et al.* 2003) for an example of a training data construction project involving bioscience text.

In using Machete, biologists will naturally generate large collections of training data for a variety of tasks of importance for text mining research. Named entity annotations, such as gene names and relations between entities are one form of output. Typed passages are another. A biologists will be able to direct the automatic highlighting sub-system by specifying that he wishes to find passages discussing the basic biology of a gene, the parameter settings for a methodology, the link between an entity and a disease, etc. Saved passages resulting from such searches will be associated with the passage. The biologist can also manually specify what type of relevance the passage has during the annotation process. Such passages and other markup valuable for text mining research will be made available to the commu-

[1] A just-in-time system works in the background automatically gathering information that will be useful in the context of what the user is currently doing. Upon request it provides this information to the user.

nity in well-organized collections of training data with little extra effort on our part.

User Acceptance

Studies of literature usage in many fields (including bioscience) demonstrate that an article is read approximately 1000 times mostly in the first five years following publication (King & Tenopir 2000). Furthermore, it is a small percentage of published articles that receive the highest percentage of readings indicating even greater overlap in the attention of a various communities. Therefore, it is almost certain that many readers of an article are reading it to satisfy similar information needs.

People with similar information needs overlap in what they annotate and the annotations of one individual are of use to others with similar information needs (O'Hara *et al*. 1998; Wolfe 2000). Highlighting, underlining, and adding margin notes on top of existing documents is pervasive in research-oriented reading and writing activities. Scientists now perform a majority of their literature search and reading activities on-line (King & Tenopir 2000; AAP 2003; King & Montgomery 2002; Marshall 2003). It is our hypothesis that the fruits of the information gathering and analysis efforts of one researcher will enable others with similar information needs to find answers more quickly.

In many small communities such as research laboratories, competitive intelligence concerns are small and individual members will be willing to share the information they have gathered. Many years of research in knowledge management indicate that even small work groups (5 to 10 people) benefit from knowledge sharing systems (Rafaeli & Ravid 2003; Cummings 2004; Allen 1977). Furthermore, KAs will serve as an important communication tool between direct collaborators, between PIs and graduate students, and between instructors and students in courses they teach. With regard to broader information sharing with the the bioscience community as a whole, we hypothesize that many biologist will be willing to share the fruits of their literature search efforts after publication of the research for which they collected the information, perhaps as supplementary material cited in the resulting article (see the last paragraph of Section).

Summary

The volume of scientific knowledge in the biological sciences in enormous and increasing rapidly. This situation demands tools that effectively organize human scientific knowledge. While useful tools such as SwissProt exist and are used by biologists, researchers continue to use the literature as their primary source of information. We hypothesize that knowledge management tools for bioscience can and should be integrated with regular literature search activity. The Machete system, now in development, helps biologists assemble answers to specific research questions from the literature and curates such answers for future use by the same or other researchers. Machete has the additional benefit that it will also provide large collections of training data for many areas of text mining research. Many studies of literature search, reading, and annotation behavior indicate that our approach is sound and may provide a means of curating human scientific knowledge without the need for full- or even part-time curators.

References

AAP. 2003. Report on the association of american publishers - professional and scholarly publishers 2003 conference. Scholarly Information Strategies, Ltd.

Allen, T. 1977. *Managing The Flow of Technology*. Cambridge, MA USA: MIT Press.

Budzik, J., and Hammond, K. J. 2000. User interactions with everyday applications as context for just-in-time information access. In *Proceedings of IUI*.

Craven, M., and Kumlien, J. 1999. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of ISMB*.

Cummings, J. 2004. Work groups, structural diversity and knowledge sharing in a global organization. *Management Science* 50(3).

Hersh, W. R., and Bhupatiraju, R. T. TREC 2004 genomics track overview. In *Proceedings of TREC*.

Hersh, W. R., and Bhupatiraju, R. T. 2004. TREC 2004 genomics track overview. In *Proceedings of TREC*.

The 2003 biocreative evaluation. www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html.

Kim, J.-D.; Ohta, T.; Teteisi, Y.; and Tsujii, J. 2003. Genia corpus - a semantically annotated corpus for biotextmining. *Bioinformatics* 19(suppl. 1).

King, D. W., and Montgomery, C. H. 2002. After migration to an electronic journal collection. *D-Lib Magazine* 8(12).

King, D. W., and Tenopir, C. 2000. *Towards Electronic Journals: Realities for Scientists, Librarians, and Publishers*, volume 22 of *Special Libraries Association*. Special Libraries Association.

Marshall, C. C. 2003. Reading and interactivity in the digital library: Creating an experience that transcends paper. In *Proceedings of the CLIR/Kanazawa Institute of Technology Roundtable*.

O'Hara, K.; Smith, F.; Newman, W.; and SelIen, A. 1998. Student readers' use of library documents: Implications for library technologies. In *Proceedings of CHI 1998*.

Rafaeli, S., and Ravid, G. 2003. Information sharing as enabler for the virtual team: an experimental approach to assessing the role of electronic mail in disintermediation. *Information Systems Journal* 13:191–206.

Rhodes, B. 2000. *Just-In-Time Information Retrieval*. Ph.D. Dissertation, MIT Media Lab.

Tanabe, L., and Wilbur, W. J. 2002. Tagging gene and protein names in biomedical text bioinformatics. *Bioinformatics* 18(8).

Wolfe, J. L. 2000. Effects of annotations on student readers and writers. In *Proceedings of Digital Libraries 2000*.