Medical Informatics Research: Summary of Breakout Group of NSF's IDM Workshop

September 14-16, 2003

Seattle, Washington

Wanda Pratt[*]

[*] Thanks to Betty Salzberg for taking extensive notes during the sessions and input from the Biomedical and Health Informatics faculty at the University of Washington.

1 – Executive Summary

In the NSF's Information and Data Management (IDM) Program Workshop of 2003, about 35 researchers gathered to discuss how medical informatics should fit as a theme for NSF-sponsored research. The gathering included principal investigators from IDM as well as a few faculty from Biomedical and Health Informatics at the University of Washington. This report summarizes our discussion in these medical informatics break-out sessions.

Most people in this group had a strong computer science background, but their specific research areas varied. Most had prior experience in medical informatics, and everyone noted that a major challenge in this type of research is to balance work that provides a direct benefit to medicine and the biomedical collaborator but also promotes scientific advancements in information and data management. Everyone expressed a desire to establish win-win collaborations with the medical community, but they were reluctant to take on projects that only served the biomedical collaborator without offering a research challenge as well.

We discussed why NSF should participate in the support of medical informatics research in terms of how both the IDM community and the biomedical community would benefit. Throughout our discussions, we focused on the key challenges that such a program should address. We came up with three key areas: ontologies, data & information integration, and knowledge discovery & analysis. By the end of the sessions, we concluded that such work requires an interdisciplinary team of biomedical experts as well as researchers from information retrieval, database systems, and human-computer interaction.

2 – Motivation

Biomedical research and clinical medicine are undergoing a revolutionary change as they use and depend on the recent explosion of readily available biomedical data, information, and knowledge. To be effective today, biomedical researchers and clinicians use large-scale accumulations of data, perform complex analyses over these data, and need efficient and effective access to the vast amount of information, particularly in the biomedical literature. Without further advancements in information and data management, this new flood of data and information will overwhelm any single biomedical researcher or clinical practitioner.

From the perspective of the information and data management community, research in medical informatics offers real-world challenges along a number of dimensions. First, the characteristics of biomedical data, information, and knowledge present a number of difficulties. Data from biomedical experiments or even medical records are usually voluminous, dirty, incomplete, and based on implicit assumptions. Information is vast and spans many media, such as text, video, and audio, often with spatiotemporal dimensions to it. Knowledge is also highly dynamic and uncertain. In addition, the environment poses significant challenges. In many cases, the users' tasks are life critical. The users are also very busy and their roles highly varied (e.g., physicians, nurses, biologists, technicians, patients, etc.) The environment also requires extreme security and privacy to protect patient-specific information. Although many other domains have portions of these characteristics, few offer the range and complexity found in the biomedical domain.

Many of the researchers in our break-out group have been frustrated by the necessity to choose between whether our research contributions are to the biomedical community or to the information and data management community. We have felt compelled to downplay the biomedical contributions in grant proposals to the NSF. Yet, in NIH proposals, we are often forced to choose a narrow, disease-specific perspective and downplay the information and data management contributions out of a fear that the proposed work will not sound feasible. Our hope is that NSF can partner with health-care agencies, such as the NIH, to foster research that makes valuable contributions to both communities.

3 – Research Challenges and Needs

In our discussions of the needs of biomedical community that intersect with the goals of the information and data management program, three key challenge areas emerged.

Ontologies

Many of us considered ontologies as a critical component for the development of effective medical informatics systems. The medical informatics community already uses many biomedical ontologies or vocabularies, such as the Unified Medical Language System (UMLS) [1] and the Gene Ontology [2], but these systems are often simplified knowledge models. They do not support anything more than isa or part-of hierarchical links among the concepts in the model. The community has developed a few complex ontologies, but their widespread use is still plagued by several significant problems. First, biomedical experts are needed to create these ontologies, but few tools support such expert development, except at the simpler level of entering specific instances for an ontology that was largely created by an experienced knowledge engineer. Second, because biomedicine is a rapidly changing field, maintenance of the ontologies becomes essential. Third, because of the life-critical nature of medical work, all aspects of the ontology must be verifiable. Thus, a tight connection between the entered knowledge and the source of that knowledge is essential. Through sophisticated processing of the biomedical literature, information retrieval research could play an important role in identifying changes in the literature and thus helping people maintain both their ontology as well the connections to the source of knowledge.

Much of the work that has been done in the area of ontologies neglects an important component to their utility, the human-computer interaction aspects. These aspects for the development, maintenance, and use of ontologies play a large role in their usefulness and adoption. With the wide variety of professional roles that users play, support for multiple perspectives and collaboration also become key. Although some researchers have made strides to address aspects of these problems, further research that integrates work from information retrieval, database systems, and human-computer interaction is desperately needed.

Data and Information Integration

Although another IDM break-out group focused on the issue of information integration, this topic came up repeatedly in our group as a key challenge as well. In terms of benefits to medical informatics, we focused on two key characteristics: diversity and documentation needs.

The biomedical domain contains the typical diversity in terms of the variety of sources of structured data that use different schemas and make different assumptions. The database community is well aware of that type of diversity and have made great strides at addressing that

problem. An additional layer of diversity comes from the type of information. Textual information, images, audio, sensor data and the more typical structured data all must be integrated to help biomedical researchers and clinicians work effectively. This type of diversity has been neglected in most of the information integration research. Thus, there is an opportunity for information retrieval researchers and database systems researchers to work together on ways to integrate such disparate types of data and information.

Because of the life-critical nature of medical work, the documentation needs for any integration process are very high. Integration systems must maintain links to the original sources of information, and provide verification that the integrity of the data has been preserved. The privacy and security regulations pose additional constraints on documentation needs and caution in the cleaning of patient of data for purposes other the authors health care.

Knowledge Discovery and Analysis

The final key challenge is in the area of knowledge discovery and analysis. Systems that would be valuable in the biomedical community must:

- Process large volumes of information
- Find useful patterns
- Detect and deal with redundancy
- Account for uncertainty
- Show justifications for discovered knowledge
- Revisit old information when we have new knowledge or information

The knowledge discovery community already explicitly investigates the ability to process large volumes of information, but relatively little work has focused on the other requirements. In addition, few researchers engage aspects of knowledge discovery from both the structured and unstructured sources. Thus, synergies between database systems and information retrieval could be fruitful. The best work would also incorporate aspects of discovery form the artificial intelligence community as well. Finally, as with the other challenge areas, human-computer interaction will be essential for producing usable knowledge discovery systems. This work needs to go beyond mere usability and incorporate new methods for visualizing the discovered knowledge that connect to how it will be used.

Summary

To meet these challenges, we need research that spans multiple areas of research. Most current research focuses on one of area exclusively, but such narrowly specialized research is unlikely to be fruitful for addressing the complex research challenges that we have identified. Teams of researchers who bring a variety of expertise in information retrieval, database systems, artificial intelligence, and human-computer interaction are needed.

4 – Recommendations

As a result of our discussions, we recommend that NSF team with NIH to jointly fund information & data management research, particularly in the areas of ontologies, data & information integration, and knowledge discovery & analysis. Other health-oriented agencies,

such as the Agency for Healthcare Research and Quality (AHRQ) and the Center for Disease Control (CDC), could also be key contributors for such research. We argue that jointly funding this research will both fuel innovations in information and data management and insure that those innovations have a direct benefit to improving health care and advancing biomedical research. Specifically, these agencies should fund research that advances research in the areas described in section 3, but also supports the pragmatic needs of the medical collaborators. NSF's contribution to such research should help insure that the work is innovative, takes a broad perspective, and applies to a large class of medical problems, rather than problems for only a specific disease. NIH's contribution is to insure that such research is grounded in real biomedical problems.

To make advances that contribute to medicine as well information and data management, such research requires funding to support the overhead of data cleaning and management necessary to make such applied research possible. Unlike the traditional research funded by the NSF, this work requires funding for programmers and support staff to manage the practical aspects, in addition to the traditional funding for students and faculty to carry out the research. The research also needs to support an interdisciplinary team including medical collaborators, who must invest their own time to insure that the research addresses real medical problems effectively, as well information and data management researchers from the information retrieval, database systems, artificial intelligence, and human-computer interaction communities. In summary, we recommend that the NSF work with other health-care agencies, such as the NIH, AHRQ, and CDC to fund innovative, long-term and large-scale research in the areas of ontologies, data & information integration, and knowledge discovery & analysis. We see these jointly funded research endeavors as critical for advancing the science of information and data management as well as health care and biomedical research.

References

[1] NLM. The UMLS Fact Sheet. 2003. http://www.nlm.nih.gov/pubs/factsheets/umls.html

[2] Gene Ontology Consortium (2001). Creating the gene ontology resource: design and implementation. Genome Res, 11(8): 1425-33.