Report on ACM Workshop on Health Information and Knowledge Management (HIKM 2006)

Li Xiong
Department of Mathematics and Computer Science
Emory University
lxiong@mathcs.emory.edu

Yuni Xia
Department of Computer Science
Indiana University and Purdue University at Indianapolis
yxia@cs.iupui.edu

1   Introduction

Health information technology is receiving a tremendous amount of attention as a strategic area that will benefit the society in the 21st century. The continued advances in healthcare such as digitization of medical records, creation of central record systems, development of healthcare data warehouses increasingly pose new challenges to information and knowledge management. The high stakes and unique characteristics of healthcare data such as the long-term value of the data, varied data quality, the complexity of the data, the privacy constraints, as well as the availability requirements in emergent situations require a special treatment of traditional information management techniques.

The inaugural Workshop on Health Information and Knowledge Management is the first in its series for presentation and exchange of research results and experiences on leading edge issues of health information and knowledge management. The mission of the workshop is to provide an open yet focused platform for researchers and practitioners from computer science, medical informatics, as well as healthcare industry to discuss current research challenges and advances and share their perspective in various aspects of health information and knowledge management. The workshop was held on November 11, 2006, in conjunction with the ACM International Conference on Information and Knowledge Management (CIKM) in Arlington, VA.

2   Workshop Themes

The workshop was very interactive, with the audience raising many questions for the speakers and a lively discussion following the technical presentations. Several overall themes emerged from the paper presentations and discussions.

Data Heterogeneity and Data Integration. An overarching complexity associated with health information is data heterogeneity. Data resides in relational databases (such as patient medical records) as well as in unstructured text forms (such as lab and pathology reports). Many healthcare information systems and warehouses need to integrate or pull data from a variety of heterogeneous data sources including electronic medical records, registry data, GenBank, etc. Data integration across the heterogeneous sources remains a challenging issue.

Medical Text Management. The high throughput and data intensive era provides enormous opportunities for decision making and clinical research. It also presents information challenges for clinicians and medical researchers to find relevant information efficiently and to obtain just-in-time information during patient encounters. New factors such as the rich context, user-oriented evaluation, and access to the data archive have to be taken into account in designing and developing information retrieval systems for health care.

Medical Data Mining. Data mining offers promising techniques in finding patterns in medical data for

public health research such as disease prediction, diagnosis, and outcomes research. However, researchers have to cope with a number of challenges in working with medical datasets as they tend to be small, high dimensional, rich in data types, and error prone. It remains an open question to have quality mining results that are easy to search and interpreted by domain experts.

Data Security and Privacy. A major barrier for integration and exchange of medical data is the privacy and security concerns. On one hand, health information need to be integrated and available for better patient care and public health research, on the other hand, health information need to be protected for confidentiality and privacy. Research in data security and privacy is of growing importance for health information. The interesting problems range from traditional ones such as access control and multi-level security to recently emerged ones such as flexible privacy preserving data mining and data sharing.

Data integrity. A major barrier in health information management is that a small but significant amount of data contains error. In some cases, it is a result of a historical rule such as using mother's Social Security Number in a child's medical record. In other cases, it is a mistake during the process of transcribing data from doctor's hand-written notes. Research advances in data verification as well as practical techniques that can be applied in health information are of particular interests.

3 Program

The workshop program includes a keynote speech and 7 paper presentations followed by a discussion. The paper presentations are divided into 3 sessions that cover a variety of topics, including medical document indexing and retrieval, medical data mining and clinical trial management, healthcare data integration and exchange, and security management.

3.1 Keynote Address

The keynote address was given by Dr. Tyrone Grandison from IBM Almaden Research with the title "Enabling the Healthcare Revolution". In his talk, he emphasized that the 21st Century has ushered in new awareness of the need to address the dire state of the Healthcare sector. There is worldwide push to leverage Information Technology to help reduce the current problems in the sector. However, many problems to be faced in this transition are either incorrectly labeled as problems or go unidentified. He gave an overview of the problems in the health care from the business, social and legal perspectives. He illustrated the technology challenges and existing solutions in various areas including modeling, standardization, storage, security and privacy, data analytics, interoperability, remote system and service science. He concluded his talk with a discussion of the areas for future exploration and a call for action. The slides are available at the HIKM workshop website[1].

3.2 Medical Document Indexing and Retrieval

Evangelos Milios first presented a paper titled "Automatic document indexing in large medical collections" [2]. He presented AMTEx, an automatic term extraction method, specifically designed for the automatic indexing of documents in large medical collections such as MEDLINE, the premier bibliographic database of the U.S. National Library of Medicine (NLM). AMTEx combines MeSH, the terminological thesaurus resource of NLM, with a well-established method for extraction of domain terms, the C/NC-value method. The performance of various AMTEx configurations in the indexing task is measured against the current state-of-the-art, the MMTx method, on a subset of MEDLINE documents. While AMTEx achieves better precision and recall than MMTx, it still suggests that term extraction in large medical document collections remains to be a challenging task.

Susan Price then presented a paper titled "Using semantic components to express clinical questions against document collections" [5]. She described a new model for describing the content of documents in domain-specific collections, using document classes and semantic components, that may supplement existing indexing and searching techniques and improve information retrieval. She also presented the results of using the model to represent clinical questions in the medical domain. They manually mapped generic questions from a clinical question taxonomy to two web-based document collections using the document classes and semantic components they identified for each collection. They successfully mapped 36 of 50 question categories in one resource, and 34 of 50 in the other. Based on the frequency of the question

---

[1]http://www.mathcs.emory.edu/hikm

types in the taxonomy, over 92% of questions were covered by the mappings in both resources.

3.3 Medical Data Mining and Applications

Carlos Ordonez (University of Houston) presented a paper titled "Comparing association rules and decision trees for disease prediction" [4]. He described a decision rule mining method in which search constraints are introduced to find only medically significant association rules and make search more efficient. Association rules are compared to predictive rules mined with decision trees on a radiology dataset. Experiments show that decision trees tend to find few simple rules, most rules have somewhat low reliability, most attribute splits are different from medically common splits, and most rules refer to very small sets of patients. In contrast, association rules generally include simpler predictive rules, they work well with user-binned attributes, rule reliability is higher and rules generally refer to larger sets of patients.

Ravi Shankar (Stanford University) presented a paper titled "Epoch: an ontological framework to support clinical trials management" [6]. The increasing complexity of clinical trials has generated an enormous requirement for knowledge and information specification at all stages of the trials, including planning, documentation, implementation, and analysis. He presented a knowledge-based framework (Epoch) to support the management of clinical trials tailored to the Immune Tolerance Network (ITN), an international research consortium developing new therapeutics in immune-mediated disorders. They currently target two areas that are vital to the successful implementation of a trial: (1) tracking study participants as they advance through the trials, and (2) tracking biological specimens as they are processed at the trial laboratories. The core of the software architecture is a suite of ontologies that conceptualizes relevant clinical trial domain.

3.4 Medical Data Integration and Exchange

Vagelis Hristidis (Florida International University) presented a paper titled "A flexible approach for electronic medical records exchange" [3]. The presented approach allows generating a customized EMR independent of existing healthcare applications and provides an on-demand, secure, efficient, and semantics-agnostic way to exchange EMRs in a collaborative environment using a declarative communication engine, called Communication Virtual Machine (CVM). CVM negotiates the capabilities of the involved parties and underlying networks to guarantee Quality of Service and presentation compatibility. It can also be customized to enforce privacy and security requirements (e.g., HIPAA) by enabling logging, authentication, and so on. A prototype of the EMR exchange approach has been implemented which integrates the i-Rounds medical record system used at Miami Children's Hospital.

Rafae Bhatti (Purdue University) presented a paper titled "Policy-based security management for federated healthcare databases (or RHIOs)" [1]. He described a context-aware policy-based security management system for health informatics. The policies are based on a set of use cases developed for the HL7 Clinical Document Architecture (CDA) standard. The system is designed to adapt well to ubiquitous healthcare services in a non-traditional, pervasive environment using the same infrastructure that enables federated healthcare management for traditional organizational boundaries. Their work also included an enforcement architecture and a demonstration prototype for the policy-based system.

Wai Gen Yee (Illinois Institute of Technology) presented a paper titled "Bridging a gap in the proposed personal health record" [7]. The emerging electronic health record infrastructure is guiding records to be stored in repositories that collectively supply a patient's comprehensive health history. However, he argued that legal and technological constraints may keep such a system from delivering health histories in a timely manner (i.e., when medical attention is needed). He presented a design for a portable personal health record system that complies with HIPAA standards of security and interaction. The authenticity of stored records on this PHR is automatically verifiable, increasing its usefulness to health care providers.

4   Final Note

There have been recent and upcoming workshops focusing on individual information management issues in general or biomedical domains that also apply to health information management, such as Workshop on Database Interoperability (InterDB)[2], Workshop

---

[2]http://www.fundp.ac.be/eco/interdb/2007/

on Privacy Data Management (PDM)[3], Workshop on Secure Data Management (SDM)[4], Workshop on Quality of Databases (QDB)[5], Workshop on Management of Uncertain Data Workshop (MUD)[6], Workshop on Data Mining in Bioinformatics (BIOKDD)[7], Workshop on Biomedical Data Engineering(BMDE)[8], and Workshop on Data Integration in the Life Sciences (DILS)[9].

The HIKM workshop was the first to bring together the individual information and knowledge management issues unique to health information. The workshop organizers and attendees envision a series of workshops building upon the success of this workshop. A proposal for a second workshop is being prepared.

5 Acknowledgements

References

[1] R. Bhatti, K. Moidu, and A. Ghafoor. Policy-based security management for federated health-care databases (or rhios). In *HIKM*, pages 41–48, 2006.

[2] A. Hliaoutakis, K. Zervanou, E. G. M. Petrakis, and E. E. Milios. Automatic document indexing in large medical collections. In *HIKM*, pages 1–8, 2006.

[3] V. Hristidis, P. J. Clarke, N. Prabakar, Y. Deng, J. A. White, and R. P. Burke. A flexible approach for electronic medical records exchange. In *HIKM*, pages 33–40, 2006.

[4] C. Ordonez. Comparing association rules and decision trees for disease prediction. In *HIKM*, pages 17–24, 2006.

[5] S. Price, L. M. L. Delcambre, and M. L. Nielsen. Using semantic components to express clinical questions against document collections. In *HIKM*, pages 9–16, 2006.

[6] R. D. Shankar, S. B. Martins, M. J. O'Connor, D. B. Parrish, and A. K. Das. Epoch: an ontological framework to support clinical trials management. In *HIKM*, pages 25–32, 2006.

[7] W. G. Yee and B. Trockman. Bridging a gap in the proposed personal health record. In *HIKM*, pages 49–56, 2006.

---

[3]http://www.ceebi.curtin.edu.au/PDM2007/
[4]http://www.hitech-projects.com/sdm-workshop/sdm07.html
[5]http://WWW.hiqiq.de/qdb/
[6]http://mud.cs.utwente.nl/
[7]http://bio.informatics.iupui.edu/biokdd07/
[8]http://www.db.is.kyushu-u.ac.jp/bmde2005/
[9]http://dils07.cis.upenn.edu/