Survey of Clinical Knowledge Management and Analysis of Unstructured Text

William Lau
HPCIO/DCB/CIT/NIH
2/3/2006
Revised 2/22/2006

1. Executive Summary

[0001] Today a large portion of biomedical information can be accessed electronically. However, research articles, technical reports, clinical notes, and many other sources of information are stored as free-form text not suited for quantitative analysis. As the amount of data grow everyday, it becomes increasingly difficult to make use of the wealth of valuable knowledge that is hidden in these documents (¶ 5). This report surveys the current capabilities and limitations of biomedical text mining from the perspective of knowledge management. Our objective is to identify areas of opportunity where our contributions can yield the highest return on investment.

[0002] Extracting and analyzing information from biomedical text is especially challenging because of the complexity and diversity of the field. Natural language processing (Section 4.1) plays a major role in text mining as it transforms text into structures that can be analyzed statistically. Many machine learning algorithms have been developed to automatically generate rules for information extraction, but the accuracy of these systems in general is still not up to par (¶ 19). We have found that tremendous progress has been made in the development of biomedical ontologies, which aims to formally define concepts and their relationships to each other in a specific domain (Section 3). Many text mining techniques have incorporated ontologies to take advantage of the existing knowledge that they provide (Section 4.2). The World Wide Web has an overwhelmingly large collection of information in free text format. The semantic Web (¶ 15), in which ontologies are indispensable resources, is an area of active research to deliver machine-understandable Web contents. We predict that Web-based applications will become more intelligent as a result. As scientists currently spend a significant part of their time to find information of *real* interest, the semantic Web effort will benefit biomedical research tremendously.

[0003] Most of text mining work done in the biomedical field has been geared toward basic research. Although tools for clinical practice have been developed, many are limited to encoding applications and other basic analyses. These systems can sometimes provide decision support by retrieving relevant information and best evidence from published literature. Physicians still need to link all the pieces together before a diagnosis can be generated. The development of case-based reasoning tools has experienced remarkable growth, but few take advantage of the rich knowledge available in electronic medical record systems (¶ 23). We believe that there is an opportunity to improve case-based decision support software that makes use of the large amount of clinical data available (¶

24-26). Healthcare professionals can draw on the results of these analyses to make more accurate diagnosis and in turn provide better treatments.

[0004] Deriving reliable knowledge requires the consideration of all the related data. Data warehouses are built in large organizations to integrate information from different independent sources (Section 5). Standardization of data exchange models makes such efforts much easier. In addition to clinical records, other data types such as images have become increasingly important in patient care (Section 6). We foresee the increase in the development of integrated systems to manage different data types. Manual annotations are usually coupled with automatically extracted features to facilitate image retrieval. The need for expressing image contents with text presents an opportunity for us to develop a text mining tool to quantify image annotations. Such work will not only improve image retrieval, but also will allow physicians to quantitatively analyze the images along with their other data.

2. Introduction to Clinical Knowledge Management

[0005] The healthcare knowledge base is expanding at an unprecedented rate. Approximately 50,000 new records are added annually to the MEDLINE database alone.[1] In addition, even though published literature has traditionally been the major source for dissemination of new scientific understanding of diseases and their management, a significant portion of existing knowledge is presented in different tacit forms, from the working knowledge of the physicians, to peer discussions, to the clinical notes in patient records.[2] The vast amount of biomedical information creates a huge challenge for most healthcare professionals to provide patients with the highest quality of care that the current knowledge base can potentially support.[3] The main problem is the lack of a framework to discover and manage the knowledge systematically. Without such a framework, it is not possible to efficiently deliver all the benefits of the insights and knowledge that have already been discovered over the years.

[0006] The relatively new discipline of *knowledge management* (KM) aims to establish an environment, utilizing information technologies, to facilitate better acquisition, generation, codification, and transfer of knowledge.[4] In the healthcare arena, a lot of effort has been focused on computerization, standardization, and automated analysis of clinical data. Many hospitals have already started to replace their paper-based record systems with computer-based technologies,[5] which support more up-to-date, easier-to-access, and more comprehensive clinical records than the traditional practice. Furthermore, the information has to be interoperable among heterogeneous databases, possibly across multiple institutions, to maximize its utility. The HL7 Clinical Document Architecture standard is a key effort in establishing a well-delineated representational model, specifying the structure of a clinical document for data exchange.[6]

[0007] In addition to interoperability, data standardization can also address semantic compatibility issues by mapping relevant text into standardized concepts in biomedical ontologies, such as the *Unified Medical Language System*[7] (UMLS). Ontologies help convey the semantics of textual information in a machine-understandable format so that

data from different sources can be reliably integrated for more sophisticated analysis.[8] As one can probably imagine, manual encoding or tagging of the entities to standardized concepts is labor intensive. Thus, various *natural language processing* (NLP) techniques have been developed or applied to make such task automatic. Several techniques and systems will be introduced later in this survey.

[0008] The standardized clinical data model provides the necessary foundation to convert data into knowledge. Relevant information can be retrieved from published best evidence to support decision making.[2] In more sophisticated applications, important connections between individual elements can be made to generate hypotheses.[1] The newly gained knowledge can assist in both diagnosis and treatments. A number of text-based data analysis methods have actually been implemented for a wide range of clinical and public health applications. For example, biosurveillance systems are used to classify patients into syndromic categories.[9] A significant increase of patients in a category may signal an outbreak of a contagious disease or a terrorist biochemical attack. These systems use fever detection algorithms for instance, to determine the presence of fever from free-text clinical reports.[10] Clinical text analysis is also useful in detecting adverse events that may be neglected from voluntary reporting. Melton *et al.*[11] developed a system to identify from discharge summaries 45 types of adverse events. The system allows healthcare institutions to learn from the events and take necessary actions to reduce the possibility of reoccurrence. These and other tools facilitate a systematic approach to manage and make use of the massive amounts of potentially useful data. Nonetheless, achieving full benefits of clinical KM will require collaborative efforts from all stakeholders, ranging from healthcare providers, to industries, to government agencies. To reach that aim, the non-profit organization OpenClinical[12] was created to promote and coordinate KM resources in patient care and clinical research.

3. Ontologies

[0009] An *ontology* links concepts to their interpretations to avoid potential ambiguity of the textual terms. For example, within the biomedical context, “cold” is a condition to describe the temperature. “Cold” is also an alternate term for “common cold”. In yet another meaning, “COLD” is an acronym for “chronic obstructive lung disease”. Furthermore, the term “heart attack” is interchangeable for “myocardial infarction”.[7] Therefore, unique identifiers are associated with each concept to handle both *polysemy* (i.e. one name with multiple meanings) and *polyonomy* (i.e. multiple names for one concept) common in natural languages.[13] If different databases tag their data elements using the same ontology, the identifiers can be used to retrieve associated data in individual databases even if the values are not exactly identical.

[0010] In addition, biomedical knowledge is often hierarchical in nature. Thus, an ontology also describes the relationships among different concepts through a set of assertions and rules.[14] Some of the common relations in a biological ontology include *part-of*, *isa*, *causes* and *prevents*.[15] The formal representation of knowledge allows a user to query a database for data that is not only directly tied to a specific term, but also uses the hierarchy in the ontology to further obtain related data.

[0011] A number of ontologies and coding systems for various biomedical domains have been developed in recent years. The UMLS developed by the US National Library of Medicine (NLM) is a consolidated repository composed of three main knowledge sources:[16] a Metathesaurus® of concepts and terms from more than 100 source vocabularies;[7] a Semantic Network consisting of various semantic types and relationships to provide a consistent categorization of every concept in the Metathesaurus; and the SPECIALIST Lexicon addressing the high degree of natural language variability of biomedical and common words in the English language. Some of the source vocabularies important to healthcare include SNOMED CT, ICD, and HL7. These are summarized below:

[0012] The *Systemized Nomenclature of Medicine Clinical Terms* (SNOMED CT)[17] is a clinical terminology developed by the College of American Pathologists. It provides comprehensive coverage of diseases, clinical findings, therapies, procedures, and outcomes. The terminology of over 366,000 concepts can be used to support recording and reporting of a patient's care in *electronic medical record* (EMR) systems. In 2003, SNOMED CT was recognized by the American National Standards Institute (ANSI) as the Health Terminology Structure Standard. The same year an agreement was reached to incorporate SNOMED CT into UMLS to encourage the use of common medical terminology in the United States.[13]

[0013] The *International Classification of Diseases* (ICD)[18] was developed by the World Health Organization to promote international comparability in the collection and organization of morbidity and mortality statistics. The first three digits of a code identify the disease type, and the fourth and fifth digits specify diagnostic subcategories. The terminology is revised periodically to reflect changes in the medical field and latest version (ICD-10) has been put in use since 1994. Although it is not intended to be a terminology for documenting clinical care, ICD has been adopted for many health management purposes. In the U.S., ICD is often used to code diagnoses on reimbursement forms submitted to insurance companies.[19]

[0014] The *Health Level 7* (HL7) group,[6] which develops the Clinical Document Architecture for data exchange, recognizes the importance of vocabulary domain specification. Such specification enhances semantic understanding by constraining the range of values that are allowed. The purpose of specifying a vocabulary domain is to ensure computability of the information in the coded fields,[20] so that for instance, a field for gender will only accept either *Male* or *Female* as its value and not something else. Its Vocabulary Technical Committee is responsible to define and maintain a vocabulary domain for each coded entry in the HL7 messages.

[0015] The Open Biological Ontologies (OBO) project is an effort to create a standard format to achieve shared use across different ontologies.[14] In recent years scientists have also investigated methods to represent existing biomedical knowledge in *semantic Web* languages, for information exchange via the Internet protocol.[15] The semantic Web,[21] which aims to deliver machine-understandable Web contents, is based on the *Resource Description Framework* (RDF). The data structure represents objects and relations in

eXtensible Markup Language (XML) format. When defining in a *Web Ontology Language* (OWL), biomedical ontologies can be made compatible with the architecture of the Web. Accessibility is enhanced when information is placed on the Web, making it easier to share and integrate knowledge from many different sources. Web-based applications will likely become more intelligent as a result of the semantic Web effort. Researchers will benefit the most as they now spend a significant part of their time to find information of *real* interest.[15]

4. Text Mining

[0016] Knowledge discovery is one of the goals in KM. A large portion of biomedical information is available in electronic format. However, research articles, technical reports, clinical notes, and many other sources of information are stored as free-form text because of the flexibility it offers.[22] Natural-language text often conveys rich concepts in a fashion that is not readily apparent.[23] The string, "Time flies like an arrow", exemplifies the complexity of natural languages. Thus, it is usually difficult to extract new knowledge from these documents. The goal of text mining is to identify non-trivial, implicit, previously unknown information in text.[22]

4.1. Natural Language Processing

[0017] Text mining has its origin from data mining.[24] However, the information in conventional data mining is usually highly structured, containing mostly numbers and symbols. Text, on the other hand, has minimal structure, governed only by grammatical rules and organizational conventions (e.g. paragraphs, indentations).[22] Arguably, text mining is a much more challenging task. It operates at various levels of granularity, including, but not limited to studying the relationships of words in a sentence, identifying the discourse of sentences, summarizing a document, and clustering documents based on their features. A majority of text mining work involves NLP because what it does essentially is to transform the text into structures that can be analyzed by data mining techniques. NLP is an interdisciplinary field where the theories of linguistics meet artificial intelligence.

[0018] The first step of NLP often entails tokenization,[25] a process to decompose the text into individual sentences, words, or perhaps morphemes. Part-of-speech tags[26] are then assigned to individual words according to their lexical classes, such as noun, verb, and adjective. The words may be transformed into their base forms called lemma. For example, 'take' is the lemma for 'takes', 'taking', 'took', and 'taken'. This process is known as lemmatization.[27] Syntactic processing involves two steps: shallow parsing and full parsing.[28] In the first step, phrasal elements are determined. Extraction of long and complex phases, for instance 'the Food and Drug Administration headquarters', can be challenging. A syntax tree is subsequently derived to identify the syntactic structure of the sentence. The interpretation of semantic content usually requires both general and domain-specific knowledge.[8] Nevertheless, ontologies can be used to tag entities belonging to certain classes in the domain (i.e. named entity recognition) and to even extract high-level relationships.

[0019] In conjunction with ontology-based approaches, rule-based techniques with rules generated either manually or automatically through induction methods have been reported.[22] Common machine learning algorithms that biomedical researchers have applied include decision trees,[29] hidden Markov models.[30] support vector machines,[31] and naïve Bayesian techniques.[32] These models are tuned by a large number of pre-classified text data, collectively called the training set. A test set of unclassified text is then used to determine the actual performance of the model. Regardless of the type of methods used, it remains difficult to develop grammar and extraction rules that yield high accuracy.[33]

4.2. Efforts in Text Mining

[0020] Although not clearly distinguished, three stages of processing can be identified in a text mining operation: *information retrieval* (IR), *information extraction* (IE), and *knowledge discovery* (KD). IR[34] aims to identify those documents that are of interest. Web search engines, e.g. Google, are the most popular applications in IR. Keyword matching, whereby each document in the search space has been indexed by a set of keywords, is the simplest technique. However, the reliability of this technique is questionable. When the intention of the query is complicated or the keywords can be manifested in more than one form, irrelevant documents are retrieved while important information can be overlooked. A better approach that most search engines utilize is the vector-space model, which associates each search term with a weight.[35] Some more sophisticated IR systems, e.g. Ask Jeeves, are able to process natural language queries to better understand users' specific intention. Some can also locate documents, potentially with the assistance of an ontology, that are relevant to the search topic even if the keywords themselves do not appear in the documents. The search engine developed by Suarez *el at*.,[36] which uses UMLS to improve search results on the Web, is an example of such a system.

[0021] Unless the objective is to identify the number of documents that contain particular keywords or character strings, the output from the IR phase is a list of documents that still requires filtering to obtain the specific information desired. IE[37] is intended to find predefined types of entities, relations or events in free text. Co-occurrence of terms is the basic means for identifying relationships. When two entities are linked by a verb, there exists a type of relationship specified by that verb. However, the task is not as easy as it may seem. Obviously, negation has to be considered. Different representations of the same entity, e.g. abbreviation and synonym, have to be taken into account as well. In fact, most systems are not good at handling anaphora resolution,[33] the process of determining to what a pronoun is referring. They are not capable of extracting relationships that extend over more than a few sentences. An ontology is a tremendous resource for solving some IE problems. In the passive use of antologies, the task is to map a term occurring in text to a concept in an ontology.[8] An example of this application is the automated encoding tool[38] that maps terms in clinical documents to UMLS codes. However, text mining aims to extract information that is novel. In ontology-driven IE, an ontology is used to actively guide and constrain analysis.[8]

[0022] Many basic tasks in IR and IE can be considered as the pre-processing stage for KD. In this stage, the text is transformed to suitable representation for analysis. Different analyses can be performed to discover new patterns and to formulate hypotheses. Methods such as classification, clustering, regression, correlation, and visualization are commonly used.[22] A number of text mining studies on biomedical literature have led to new scientific discoveries.[39] A classic example of such was the Swanson study.[40] The study connected seemingly unrelated articles in the literature and discovered the relation between Raynauds disease and fatty acids in fish oil. The finding was later verified by other scientific studies.[39] Researchers have also recently explored the use of ontologies to find new knowledge from text. For example, using the UMLS ontology, the natural-language processor, MEDSYNDIKATE,[41] is able to handle various discourse structure phenomena and extract knowledge from medical finding reports.

[0023] Most of text mining work in the biomedical field is geared toward basic research, primarily on genomic analysis. Although tools for clinical practice have been developed, many are limited to encoding applications and other basic analyses. A number of projects discussed in this survey have suggested the tremendous potential for text mining to improve diagnoses and treatments. An area of development where text mining can bring contributions is case-based reasoning[42] (CBR). In accordance with evidence-based medicine, the CBR approach to solving new problems is based on the solutions of similar problems in the past. In medicine where decisions can mean life or death, it is especially critical to take into account past experience beyond the general knowledge found in textbooks.

[0024] Although the use of CBR tools has experienced remarkable growth, the wealth of knowledge hidden in the free text of EMRs remains largely untapped.[43, 44] If such knowledge can be capitalized, more advanced case-based decision support software can be implemented. In addition to retrieving similar cases, recommendations can be generated based on analysis on the EMRs. The value of such software can be demonstrated in following hypothetical case study:

[0025] A text mining system has been developed to identify and encode the symptoms, diagnoses, treatments, and outcomes of patient records as they are entered into the database. A patient comes to the emergency room with a set of serious but rare symptoms. A diagnosis cannot be made based on the patient's medical history. Various tests have been run, but the physicians cannot wait too long for the test results to come back because the patient's condition continues to deteriorate. The hospital is equipped with the text mining system, and in a fraction of time it retrieves more than 50 similar cases from around the world. The physicians determine that the patient has contracted a deadly virus usually found in Asia. Instead of having to go through all of the cases, the physicians are presented with a summary showing what the possible treatments are and their associated outcomes. They agree on a therapy that is the most appropriate for the current situation. Because of the timely intervention, the patient's life is saved.

[0026] This case study shows that advanced case-based decision support systems can reduce the time physicians take to link individual elements together, allowing them to optimally solve problems in time-constrained circumstances. It can significantly benefit care providers who may not have experience with the presented symptoms and can prevent them from overlooking certain aspect of the case. In this case of decision support, the implicit knowledge encoded in the medical records may not necessarily be transformed into explicit knowledge. Human judgment is still the main component of the decision making process.

5. Data Warehousing

[0027] For large decentralized organizations, such as the National Institutes of Health (NIH), there are usually a large number of independent databases across various labs and departments. A data warehouse extracts important data from production systems according to a schema.[45] It provides a consolidated view of the organization's data to support queries concerning best practices, operational effectiveness, and cost efficiency.[46] The warehouse refreshes its data periodically, ideally when the workload of the systems is low to minimize performance impacts. In addition, clients of the data warehouse do not have direct access to the data sources, ensuring data integrity of the operational databases.[47]

[0028] Most data warehousing schemes extract and transform data from many local data sources and place them into a physically separate repository for online access, reporting, and data mining.[45] The task usually requires considerable effort in consolidating inconsistent data into one coherent set.[47] The standardization of semantics and data-exchange models reduces development time and helps ensure data quality.[48] However, certain clinical applications are still limited from this integrated approach because the data may not be up-to-date since last being synchronized with the sources.

[0029] In addition to aggregating similar data to provide centralized access, composition of complementary data from multiple sources may lead to discovery of new knowledge that may be difficult to derive from looking at the data separately. For example, a system supporting complex queries of a patient's genomic, proteomic, and clinical data facilitates a comprehensive analysis of all information. The detailed knowledge about the patient allows for personalized therapy, which can lead to improved outcomes for the patient.[49] It may, however, be difficult to have a data exchange standard that works across such diverse domains. The *federated* data warehouse approach[45] forms a hierarchy of warehouses that allows flexibility at the local level and also maintains a view from the global perspective. The local warehouses in the federation handle data within a specific domain whereas global warehouses store common data. Common data can flow downward from these repositories, while important summaries flow upward from the bottom of the hierarchy.

[0030] Data warehousing projects have been conducted at the University of Michigan Health System and the Ohio State University Medical Center since 1998 and 2003, respectively.[46, 50] The warehouses consolidate clinical, financial, employee records as

well as information from the literature and other external sources. The projects provide a positive return of investment from the knowledge that can be derived from the data warehouses and the application of that knowledge to improve communication, education, research, and patient care. The success of these projects lies on the convergence of people, process, data, and technology. Although both groups have to face a number of difficult challenges and failures, the projects have brought about a more robust information management infrastructure to manage their ever expanding knowledge assets.

6. Analogy to Image Data Management

[0031] The medical imaging community is confronted with many barriers in image management similar to the text KM challenges discussed thus far. In the 2004 JASON report,[51] some of the imaging issues being identified were standardization, information extraction, and data archiving. These issues arise partly because of the proliferation of the number of digital medical images, making the management and access to large image repositories increasingly difficult.[52]

[0032] With the *digital imaging and communications in medicine* (DICOM)[53] standard, images can be exchanged between systems, irrespective of manufacturer. In addition, each DICOM file has a header where patient information along with other specifications can be stored. It is not unusual that radiologists will compare a set of images to identify abnormalities or to evaluate treatment progress. The lack in standardization is a common set of metrics for comparing similarities between different images. Calibration, intrinsic uncertainty due to the limitation of the modality, and subjective judgment all contribute to significant variations in interpretation of imaging data.[51] Computer-assisted quantitative analysis can provide sets of parameters, enhancing the objectiveness of image comparison.

[0033] Various techniques have been developed to extract information in images. However, most *content-based image retrieval* (CBIR) systems only deal with low-level features, such as color, size and shape.[52] Other than object recognition,[49, 54] it remains impossible to extract high-level semantics from images. Manual annotations are usually coupled with automatically extracted features to facilitate image retrieval.[55] A few indexing systems have been proposed to extract UMLS concepts from free text in radiology reports.[56, 57] The next step will be to develop a text mining tool to quantify image annotations, such as the size and location of a tumor. Such work will not only improve image retrieval, but will also allow physicians to quantitatively analyze the images along with other data. In addition to feature-based techniques, many systems locate and cluster images based on the assumption that there exists a correlation between the characteristics of an image and the concepts it depicts.[52] Some of these systems use methods that are commonly employed in text retrieval (such as the *inverted file method,*[58] which restrict the subspace spanned by the search.)

7. Conclusion

[0034] Clinical KM is more than just technology, since it changes the way knowledge is managed. Advances in information technology have enabled systematic processing of institutional data, information and knowledge. The shift from paper-based to computer-based processing has opened up an opportunity for making better use of the available information. Through coordinated KM, the information stored in EMRs will no longer be used for administrative purposes only, but also for decision support as well as for clinical research. Healthcare systems and biomedical knowledge repositories are being integrated to form inter-departmental and inter-institutional networks. The vision is to turn tacit and implicit knowledge into explicit knowledge that healthcare professionals can access whenever and wherever they need it.

[0035] A majority of the knowledge generated every day is hidden in text. Natural language processing techniques are being used for many purposes, from document retrieval, to information extraction, to automatically encoding. Computers will undoubtedly play an increasing role in knowledge discovery. To alleviate the problem of information overload, they have to run more analyses previously done by people and present users only with the most relevant information. A requisite for achieving this goal is that computers have to be able to understand the concepts in text. Ontologies are being developed to formalize the representation of knowledge within a specific domain. In an ontology, each concept is unambiguously defined and their relationship with one another is established explicitly. Isolated islands of knowledge also have to be integrated to derive new knowledge. Therefore, systems must be able to exchange data efficiently. Through several initiatives, such as the semantic Web and the Clinical Document Architecture, standard communication protocols are formed. Data warehouses are also created in large organizations to facilitate best practices and foster collaborations.

[0036] Information is increasingly presented in different data types other than text. Images share a lot of similarities with text. CBIR systems are developed to address some of issues in image management. Some of these systems use annotations to supplement visual parameters extracted from images automatically. If enough resources are placed in this area, it may not be long until case-based decision support tools are integrated with both CBIR and EMR systems. When this is realized, an oncologist can ask the system a question like “For all patients in the United States who have reported symptoms similar to those in this case, give me the records of those patients whose brain MRI scans subsequently indicate the presence of a tumor in the visual cortex.” This query may require a lot of computational power, but high performance computing can make real-time processing possible. The wealth of information provided to the physicians greatly enhances the quality of their diagnoses and treatments.

[0037] This survey has identified several areas of opportunities where more research and development can lead to significant improvements in KM:

- Incorporating NLP into CBR systems to interpret textual information in EMRs and support clinical decision making.
- Qualifying image annotations to facilitate CBIR and objective image comparison.

- Implementing data warehouses with abilities to effectively manage and retrieve information from different data types, including text and images.

[0038] Technology itself cannot realize all the potential benefits that can be supported by knowledge scattered across disciplines. People, processes, data and technology are matching pieces of the puzzle in KM. However difficult it may be, the four essential components have to work together so that the acquisition, generation, codification, and transfer of knowledge can be carried out in the most effective way. The consequence of effective clinical KM is better scientific research and healthcare beyond the limits of any individual or even institution can achieve alone.

8. References

1. Cohen, A. M. & Hersh, W. R. A survey of current work in biomedical text mining. *Brief Bioinform* **6**, 57-71 (2005).

2. Abidi, S. S., Cheah, Y. N. & Curran, J. A knowledge creation info-structure to acquire and crystallize the tacit knowledge of health-care experts. *IEEE Trans. Inf. Technol. Biomed.* **9**, 193-204 (2005).

3. OpenClinical. The medical knowledge crisis and its solution through knowledge management. (2000).

4. Bali, R. K., Feng, D. D., Burstein, F. & Dwivedi, A. N. Introduction to the special issue on advances in clinical and health-care knowledge management. *IEEE Trans. Inf. Technol. Biomed.* **9**, 157-161 (2005).

5. Lorence, D. P. & Churchill, R. Clinical knowledge management using computerized patient record systems: is the current infrastructure adequate? *IEEE Trans. Inf. Technol. Biomed.* **9**, 283-288 (2005).

6. http://www.hl7.org/.

7. National Library of Medicine. UMLS Knowledge Sources 2005AC Documentation. (2005).

8. Spasic, I., Ananiadou, S., McNaught, J. & Kumar, A. Text mining and ontologies in biomedicine: making sense of raw text. *Brief Bioinform* **6**, 239-251 (2005).

9. Tsui, F. C. *et al*. Technical description of RODS: a real-time public health surveillance system. *J. Am. Med. Inform. Assoc.* **10**, 399-408 (2003).

10. Chapman, W. W., Dowling, J. N. & Wagner, M. M. Fever detection from free-text clinical records for biosurveillance. *J. Biomed. Inform.* **37**, 120-127 (2004).

11. Melton, G. B. & Hripcsak, G. Automated detection of adverse events using natural language processing of discharge summaries. *J. Am. Med. Inform. Assoc.* **12**, 448-457 (2005).

12. http://www.openclinical.org.

13. Fung, K. W. *et al*. Integrating SNOMED CT into the UMLS: An Exploration of Different Views of Synonymy and Quality of Editing. *J. Am. Med. Inform. Assoc.* **12**, 486-494 (2005).

14. Bard, J. B. L. & Rhee, S. Y. Ontologies in biology: design, applications and future challenges. *Nature Reviews Genetics* **5**, 213-222 (2004).

15. Mukherjea, S. Information retrieval and knowledge discovery utilising a biomedical Semantic Web. *Brief Bioinform* **6**, 252-262 (2005).

16. http://www.nlm.nih.gov/mesh/umlsforelis.html.

17. http://www.snomed.org/snomedct/index.html.

18. http://www.who.int/classifications/icd/en/.

19. Alexander, S., Conner, T. & Slaughter, T. Overview of inpatient coding. *Am. J. Health. Syst. Pharm.* **60**, S11-14 (2003).

20. Bakken, S., Campbell, K. E., Cimino, J. J., Huff, S. M. & Hammond, W. E. Toward vocabulary domain specifications for health level 7-coded data elements. *J. Am. Med. Inform. Assoc.* **7**, 333-342 (2000).

21. http://www.w3.org/2001/sw/.

22. Natarajan, J., Berrar, D., Hack, C. J. & Dubitzky, W. Knowledge discovery in biology and biotechnology texts: a review of techniques, evaluation strategies, and applications. *Crit. Rev. Biotechnol.* **25**, 31-52 (2005).

23. Nasukawa, T. & Nagano, T. Text analysis and knowledge mining system. *IMB Systems Journal* **40**, 967-984 (2001).

24. Hearst, M. A. *Untangling Text Data Mining* (Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, ACM, 1999).

25. Webster, J. J. &Kit, C. ***Tokenization as the initial phase in NLP*** (Proceedings of the 14th conference on Computational linguistics Ser. 4, Association for Computational Linguistics, Morristown, NJ, USA, 1992).

26. Manning, C. D. & Schutze, H. in *Foundations of Statistical Natural Language Processing* 341-380 (The MIT Press, Cambridge, MA, 1999).

27. Plisson, J., Lavrac, N. & Mladenic, D. *A rule based approach to word lemmatization* (Proceedings of the 7th International multi-conference Information Society, 2004).

28. Jurafsky, D. & Martin, J. H. in *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition.* (Prentice Hall, Upper Saddle River, NJ, 2000).

29. Camoglu, O., Can, T., Singh, A. K. & Wang, Y. F. Decision tree based information integration for automated protein classification. *J. Bioinform Comput. Biol.* **3**, 717-742 (2005).

30. Yeganova, L., Smith, L. & Wilbur, W. J. Identification of related gene/protein names based on an HMM of name variations. *Comput. Biol. Chem.* **28**, 97-107 (2004).

31. Yabuki, Y., Muramatsu, T., Hirokawa, T., Mukai, H. & Suwa, M. GRIFFIN: a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model. *Nucleic Acids Res.* **33**, W148-53 (2005).

32. Deng, X., Geng, H. & Ali, H. Learning Yeast Gene Functions from Heterogeneous Sources of Data Using Hybrid Weighted Bayesian Networks. *Proc. IEEE Comput. Syst. Bioinform Conf.* **4**, 25-34 (2005).

33. Jensen, L. J., Saric, J. & Bork, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.* **7**, 119-129 (2006).

34. Beaza-Yates, R. & Ribeiro-Neto, B. in *Modern Information Retrieval* (ACM Press, New York, 1999).

35. Yandell, M. D. & Majoros, W. H. GENOMICS AND NATURAL LANGUAGE PROCESSING. *Nat. Rev. Genet.* **3**, 601-610 (2002).

36. Suarez, H. H., Hao, X. & Chang, I. F. Searching for information on the Internet using the UMLS and Medical World Search. *Proc. AMIA. Annu. Fall. Symp.*, 824-828 (1997).

37. Hobbs, J. *The generic information extraction system* (Proceedings of the Fifth Message Understanding Conference (MUC5), Morgan Kaufmann Publishers, Inc., San Francisco, CA, 1993).

38. Friedman, C., Shagina, L., Lussier, Y. & Hripcsak, G. Automated encoding of clinical documents based on natural language processing. *J. Am. Med. Inform. Assoc.* **11**, 392-402 (2004).

39. Swanson, D. R. Medical literature as a potential source of new knowledge. *Bull. Med. Libr. Assoc.* **78**, 29-37 (1990).

40. Swanson, D. R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.* **30**, 7-18 (1986).

41. Hahn, U., Romacker, M. & Schulz, S. Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. *Pac. Symp. Biocomput.*, 338-349 (2002).

42. Kolodner, J. L. E. -. An introduction to case-based reasoning. *Artif. Intell. Rev.* **6**, 3-34 (1992).

43. Pantazi, S. V., Arocha, J. F. & Moehr, J. R. Case-based medical informatics. *BMC Med. Inform. Decis. Mak.* **4**, 19 (2004).

44. Bichindaritz, I. & Marling, C. Case-based reasoning in the health sciences: What's next? *Artif. Intell. Med.* (2006).

45. Kerschberg, L. E. -. in *Knowledge Management in Heterogeneous Data Warehouse Environments* 1, 2001).

46. DeWitt, J. G. & Hampton, P. M. Development of a Data Warehouse at an Academic Health System: Knowing a Place for the First Time. *Acad. Med.* **80**, 1019-1025 (2005).

47. Kerkri, E. M. *et al*. An Approach for Integrating Heterogeneous Information Sources in a Medical Data Warehouse. *J. Med. Syst.* **25**, 167-176 (2001).

48. Sujansky, W. Heterogeneous Database Integration in Biomedicine. *J. Biomed. Inform.* **34**, 285-298 (2001).

49. Hu, H. *et al*. Biomedical informatics: development of a comprehensive data warehouse for clinical and genomic breast cancer research. *Pharmacogenomics* **5**, 933-941 (2004).

50. Cain, T. J., Rodman, R. L., Sanfilippo, F. & Kroll, S. M. Managing knowledge and technology to foster innovation at the Ohio State University Medical Center. *Acad. Med.* **80**, 1026-1031 (2005).

51. Stubbs, C. *et al*. The Computational Challenges of Medical Imaging. **JSR-03-300** (2004).

52. Muller, H., Michoux, N., Bandon, D. & Geissbuhler, A. A review of content-based image retrieval systems in medical applications--clinical benefits and future directions. *Int. J. Med. Inf.* **73**, 1-23 (2004).

53. Graham, R. N. J., Perriss, R. W. & Scarsbrook, A. F. DICOM demystified: A review of digital file formats and their use in radiological practice. *Clin. Radiol.* **60**, 1133-1140 (2005).

54. Perronnin, F., Dugelay, J. L. & Rose, K. A probabilistic model of face mapping with local transformations and its application to person recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1157-1171 (2005).

55. Pfund, T. &Marchand-Maillet, S. *Dynamic multimedia annotation tool* (Proc. for SPIE Internet Imaging III Ser. 4672, SPIE, 2001).

56. Lowe, H. J., Antipov, I., Hersh, W., Smith, C. A. & Mailhot, M. Automated semantic indexing of imaging reports to support retrieval of medical images in the multimedia electronic medical record. *Methods Inf. Med.* **38**, 303-307 (1999).

57. Hersh, W., Mailhot, M., Arnott-Smith, C. & Lowe, H. Selective automated indexing of findings and diagnoses in radiology reports. *J. Biomed. Inform.* **34**, 262-273 (2001).

58. Squire, D. M., Muller, W., Muller, H. & Pun, T. Content-based query of image databases: inspirations from text retrieval. *Pattern Recog. Lett.* **21**, 1193-1198 (2000).