S. Chu

The University of Auckland,
New Zealand

Synopsis

Information Retrieval and Health/Clinical Management

Introduction

The healthcare industry is one of the most, if not the most, information intensive industries. Medical information/knowledge and clinical data are growing at close to explosive rates each day. Ten years ago, medical publications were added to the world's biomedical journal collections at the rate of approximately 3,000 per month. Today, the volume of bibliographic citations is growing at 1,000 per day in Medline alone [1]. Hospitals also generate huge amounts of healthcare data. It has been estimated that an acute care hospital may generate up to five terabytes of data a year [2]. The volume of information and knowledge available today has far exceeded the cognitive capability of the human brain. However, availability and use of accurate information/knowledge are crucial for delivery of quality healthcare to consumers. The demand by society and professional organizations on use of evidence-based practice to help improve quality of care also adds great pressure on healthcare professionals to regularly access best quality information and knowledge during the processes of healthcare planning, decision and delivery.

Computer-assisted information retrieval and processing provides effective mechanisms to rapidly retrieve and present the relevant information/ knowledge required for supporting quality decision-making and help overcome the human cognitive constraints. Biomedical publications and considerable volumes of clinical data are created and stored as free text documents. However, numerical machines, such as computers, are not designed to process free text effectively and structured query methods, such as SQL (structured query language) cannot be easily employed to manipulate free texts.

The benefits of being able to effectively process free text medical documents and retrieve from them relevant information and knowledge are substantial to meet the needs of healthcare professionals. However, significant technical hurdles still exist. The papers included in the 2002 IMIA Yearbook section on 'Health and clinical management' have examined and proposed some of the techniques of free text processing as useful solutions for health information retrieval. Based on these selected papers, I have categorized the issues related to healthcare information retrieval into those pertinent to (a) searching and retrieval techniques, (b) intelligent information presentation, (c) information retrieval and delivery on the Internet. Within the constraints set for this paper, I have attempted to review each issue as widely as possible and have included some discussions on future directions and developments.

Free Text Searches and Information Retrieval

1 Basic Text Searching by Pattern Matching

A number of text search techniques are available. Major categories commonly involve the application of exact string pattern matching algorithms. In most cases, they fall into the 'brute-force' (or naïve) implementation. The text string to be searched is matched to the texts in the target documents using 'character comparison and slide to next comparisons' approach. It is possible to optimize the pattern matching algorithms to produce quite fast search results [3].

Basic free text searches, either in the form of single keyword searches or text phrase searches, have the major advantage of not requiring the source texts to be pre-processed for fast queries. Commonly used text search/matching algorithms, including the Boyer-Moore-Horspool, have been evaluated by a number of experimental studies as very efficient string matching algorithms [3-5]. These algorithms can play an important role in supporting dynamic searches and retrieval of free text medical data, including clinical progress notes, pathology and radiology reports, and medical imaging reports stored in electronic medical records in which pre-processing is either difficult or not economical.

When utilizing any free text search techniques, the user needs to choose between exact text phrase matching or expanded search (i.e. combining search terms in a text phrase with operators, such as 'and', 'or') strategies. Using the exact match strategy has the benefit of returning documents more relevant to user requirements, but may also face the risk of quite few or zero document hits. The expanded search strategy, on the other hand, can bring back more documents, but has the disadvantage of returning excessively large document sets with varying degrees of relevancy to user requirements.

Morphological differences between the texts in the target documents and the search words used often further hamper the search results [6, 7], and are beyond the technological spectrum of basic free text searching/pattern matching techniques. This issue is particularly problematic in healthcare, in which the biomedical language is packed with many interchangeable terms, such as common cold and coryza, chest pain and angina pectoris, fever and pyrexia, lock-jaw and trismus, arrhythmias and dysrhythmias (although strictly speaking these two terms have quite different meanings), weakness and paresis, and many others. Substantial user frustration when clinicians attempt to search and retrieve free text information from electronic medical records can arise.

2. Indexing and use of Dictionaries/ Thesauri in text searching

To rectify some of the shortcomings of basic free text searching/pattern matching techniques, some form of pre-processing of both the search terms and the target text documents is considered a desirable approach. A number of studies had confirmed the use of indexing systems as a highly effective technique to reduce the search space and query time, hence producing very fast and accurate

searches on the rapidly growing free text document databases [8, 9]. While pattern matching style of free-text searching provide the benefit of search flexibility, search speed often suffers as the document store grows. Search speed and accuracy could be greatly enhanced by the use of indexing systems based on some standard descriptors or dictionaries. Using search terms generated from standard dictionaries also helps resolve the morphological differences problems and thus reduces user frustrations by minimizing the rates of missed-hits/failed searches. The effectiveness of this combined search approach has been confirmed by repeated studies since the early 1970s [10-13].

It is possible to build indexing systems (based on some form of standard terminology/dictionary) that are simultaneously sublinear in search space overhead and in query time [14]. Processing speed can be further enhanced by loading the entire list of indices (or its address space) onto the secondary memory of the document database server. The construction of a standard dictionary, however, presents considerable challenges. For example, the need to resolve morphological differences exists not only between authors and retrievers, but also morphological problems within the biomedical language itself. The biomedical language is loaded with morphologically similar (e.g. leucocytes, leukaemia) and semantically ambiguous (e.g. diaphoresis and diaphysis) morphemes. The shear size of biomedical terminologies and their rapid growth and dynamic nature also make the maintenance of such dictionaries extremely difficult, if not at all impossible.

The morphological segmentation, together with the automated segmentation engine developed by Schulz and Hahn [15], represented an effective interactive tool to automate the pro-

cesses of identifying and linking the semantically identical morphemes. This tool can be trained to automatically segment words imported from source texts and then link semantically identical morphemes to construct a thesaurus/ repository of morphologically and semantically related biomedical terms. The technique also provides a solution to facilitate easy, dynamic updates of the dictionary in the environment of rapidly growing biomedical terminologies. The dictionary created can be used to index target free-text documents (publications and electronic healthcare record documents) and to facilitate rapid, more effective searches and information retrieval by users.

Intelligent Information/ Knowledge Search and Presentation

The conventional method of text searching and retrieval returns free text document sets (matching the search terms) to users with little regard to the context of the search terms within the document and the relevancy/precision of the documents to the users. Low relevancy of document sets is a major cause of user dissatisfaction and abandonment of search efforts. Users' inexperience in determining appropriate search terms and strategies often aggravate the stress and add much frustration to the search processes.

Interactive graphical user interfaces today have replaced text-based search screen design as standard interfaces in almost all information retrieval engines, especially those used in libraries. Significant amount of 'smartness' can be built into the graphical interface to provide such functions, like clear display of search descriptors from standard dictionary for users to choose from, visualization of the search query building process, and guiding the user's query modification process through the

display of search results or number of hits. In the absence of human information search experts, graphical display of the information with automated query guides as a built-in search engine function produces far better results than basic free text searches that generate thousands of non-specific hits [16, 17].

A simply strategy to improve search result quality is the application of some form of information ranking techniques [18]. This approach typically requires the use of knowledge-bases containing a set of relevance indices and their relative ranking weightings, and an inference engine to manipulate the rules and the values of the ranking indices. The documents retrieved are assigned a relevance value by the inference engine based on the locations and frequencies of the search terms within the documents.

Another approach highly relevant to health/clinical management is the utilization of a categorization tool that dynamically groups and displays the document sets returned from a query under a number of specific and meaningful headings [19]. For example, a query on the concept 'breast cancer' will return a document set that can be dynamically categorized and displayed under the groups of causes, diagnosis, treatment, complications of diagnosis/ treatment, and prognosis, etc. Graphically displayed, the categorized search result becomes contextually meaningful to the user.

It is technically trivial to graphically present search terms (and their semantically related morphemes) from a biomedical dictionary in a tree-like structure. Users can then simply point-and-click to select and combine relevant terms and visualize/refine the construction of the Boolean queries [10]. If such improved search engine can be further enriched by adding a dynamic categorization tool, like the one reported in [19], the quality of search outputs (and hence their usefulness) could be dramatically enhanced.

Information Retrieval and Delivery on the Internet

Explosive developments in the Internet have made Web-based health documents increasingly important resources not only for healthcare professionals, but also for health consumers. The adoption of hypertext markup language (HTML) and extensible markup language (XML) as the de facto standards facilitates easy and rapid development of information delivery and sharing applications.

The use of free text search techniques on HTML pages to perform biomedical literature searches on remote hypertext transport protocol (HTTP) servers is a common practice in nearly all healthcare and education facilities today. It is also possible to search on, and retrieve, medical images using free text search terms submitted via an HTTP form [20]. Intelligent retrieval engine receives the free text search terms from a Web browser, map the search term to a standard term for medical image identification and use the standard term to retrieve the relevant medical image(s). The same technique can be applied to retrieving free text clinical progress notes, pathology and radiological reports.

Rapid advances in, and adoption of, Internet information access/distribution technology raise a number of highly important issues. These issues, especially those related to security and confidentiality, urgently need to be addressed. As the focus of this synopsis is not on the Internet, these issues are only touched upon briefly.

(a) Security and Confidentiality

The Internet has been designed as an open system. As such security is its weakest link. Technological capabilities (especially in the area of security) of many medium- to small- healthcare facilities is still relatively rudimentary. Technology, such as public key infrastructure (PKI), has been considered by many, particularly small clinics, as too expensive and too complicated. The use of a smart card based token (randomly generated security key) combined with user-selected password to produce a use-once only access code is a highly secure, and simple and affordable, authentication solution for all healthcare facilities, big or small. De-identification of clinical data for non clinical use and application of encryption technologies [21], such as the 3-DES (Data Encryption Standard defined encryption technology), are also crucial data protection measures which should not be ignored.

(b) Distributed Data Storage and Access Speed

The volume of Web-based documents is increasing rapidly. They are also distributed across disparate data storage and network systems. Accessing these documents from the distributed sources presents significant technical challenges. A middleware-based directory service offers the most reliable and cost effective solution for multi-organisational/regional information access. But even if the documents are located quickly, pulling large document sets and medical images across the network can create serious network traffic problems necessitating serious planning and considerable investments in network infrastructures.

(c) Reliability Issues

The Internet is essentially an unregulated (or more likely unregulable) environment. Quality of information published on Internet varies widely depending on the information sources.

Some information is potentially harmful rather than beneficial to the consumers. This can post significant risks to the less knowledgeable. One possible solution is the establishment of some international rating organization responsible for the establishment of health information evaluation criteria and publication of ratings of health-related Web sites based on the validated evaluation rules. However, rapid increases and collapses of Web sites make such activity extremely difficult and resources intensive.

Future Directions and Development

Free text biomedical documents contain huge volumes of knowledge embedded within the jungle of unstructured texts constructed by many authors. Conventional retrieval techniques (including those with categorization functions) recall abstracts and documents in their entirety, without ability to process and present the knowledge embedded within the text structures. Knowledge-based information retrieval systems focusing on knowledge acquisition from the documents, knowledge representation, knowledge-base refinement and knowledge retrieval mechanisms are required to fill this rapidly increasing demand.

Research on biomedical knowledge extraction (from medical texts), representation and retrieval using techniques, such as conceptual graph (CG) formalisms has produced promising results [22]. Canonical conceptual graphs are formed with biomedical concepts interlinked by semantic relations. Using graph construction rules, these canonical graphs may be combined to derive new CG that build up an entire sentence structure, and eventually the whole document. CG applications have been used in automated coding of clinical documents, such as patient discharge summaries, pathology and radiology reports [23, 24]. Relations between interventions and outcomes (clinical effectiveness) knowledge can be adequately represented by CG and the coding rules. Efficient clustering (for clustering related CG documents) and graph query techniques have been developed, thus allowing rapid knowledge retrieval from the conceptual graph document sets [25]. The remaining challenges include optimization of the graph query techniques and the development of integrated applications/software modules for information/knowledge retrieval and clinical information management systems.

Query agents have been/are being developed to assist users in generating suitable information retrieval queries, and in learning to adjust the queries according to user demand so that the document sets returned are contextually appropriate and best suited to user requirements [26]. The ultimate goal will be the creation of intelligent query agents that can automatically roam the Internet information/knowledge spaces and bring back knowledge relevant to the user based on user's patient profile or clinical problems. Agents could also be designed to poll/mine the clinical databases for relevant intervention-outcome data relations and submit the knowledge for automated clinical guideline improvements.

Conclusion

The papers selected for this section represent some excellent work done in various areas of information retrieval. Achievements in document/information retrieval in the past have contributed significantly to enhancing the use of information in health and clinical care management. The explosive growth in biomedical literature and pressure for knowledge retrieval have set challenging targets for the health informatics community. Already research in knowledge extraction, representation and retrieval using conceptual graphs and intelligent agents has yielded exciting results. The challenge is to further improve these technologies and to build them into integrated software applications to support automated knowledge retrieval relevant to user's patient problem/profile.

References

1. http://www.nlm.nih/gov/pubs/factsheets/online_databases.html
2. Crowe BL, McDonald JG. Evaluation of developments in storage and retrieval systems for health information systems. Proc HIC'99 – 7th National Health Informatics Conference, Hobart, Australia, 28-31 August 1999.
3. Berry D. Combing Boyer-Moore string search with regular expressions. C/C++ User Journal 2000 June;18(6):32-7.
4. De Moura SE, Navarro G, Ziviani N, Baeza-Yates R. Fast and flexible word searching on compressed text. ACM Transactions on Information Systems 2000 April;18(2):113-39.
5. Lovis C, Baud RH. Fast exact string pattern-matching algorithms adapted to the characteristics of the medical language. J Am Med Inform Assoc 2000 August; 7(4):378-91.
6. Wenzel F. Solution of morphological problems in free text retrieval during segmentation. Nachrichten für Dokumentation 1979 Dec.;30(6):212-8.
7. Doszhocs TE. Implementing an associative search interface in a large online bibliographic database environment. Proc the 39th FID Congress; 1980. p. 295-7.
8. Asokan N, Ranka S, Frieder O. A parallel free-text searching system with indexing. Proc International Conference on Databases, Parallel Architecture and their Applications, LA, USA; 1990. p. 519-21.
9. Cutting D, Pedersen J. Optimizations for dynamic inverted index maintenance. Proc the 13th International Conference on Research and Development in Information Retrieval, NY, USA; 1990. p. 405-12.
10. Xia L. Designing a visual interface for online searching. Proc The 62nd ASIS Annual Meeting NJ, USA; 1999. p.390-5.
11. Ojala M. Research into full-text retrieval. Database 1990 August;13(4):78-80.
12. Henzler RG. Free or controlled vocabularies: some statistical user-oriented evaluation of

biomedical information. International Classification 1978 March;5(1):21-6.

13. King DW, Neel PW, Wood BL. Comparative evaluation of the retrieval effectiveness of descriptor and free text search systems using CIRCOL. Research Report, Westat Research Inc., Rockville, MD, USA; January 1972.

14. Baeza-Yates R, Navarro G. Block addressing indices for approximate text retrieval. J Am Soc Inf Sci 2000 January;51 (1):69-82.

15. Schulz S, Hahn U. Morpheme-based, cross-lingual indexing for medical document retrieval. Int J Med Inf 2000; 8-59:87-9.

16. Faragher J. Questions, questions, question and answer software. Information Age 2001 January; 27-31.

17. Skolnick T. Stack search: a graphical search model. Proc International Conference on Intelligent User Interfaces NY, USA; 1999. p.190-4.

18. Lee J, Grossman D, Frieder O, McCabe MC. Integrating structured data and text: a multi-dimensional approach. Proc International Conference on Information Technology: Coding and Computing NV, USA; March 2000. p.27-9.

19. Pratt W, Fagan L. The usefulness of dynamically categorizing search results. J Am Med Inform Assoc 2000 December;7 (6):605-17.

20. Tang YK, Chiang TT. Intelligent retrieval of medical images from the Internet. Proc of Spie: The International Society for Optical Engineering, USA 1996;2711:440-8.

21. http://www.viacorp.com/crypto.html

22. Volot F, Joubert M, Fieschi M. Review of biomedical knowledge and data representation with conceptual graphs. Methods Inf Med 1998 January;37(1):86-96.

23. Delamarre D, Burgun A, Seka LP, Le Beux P. Automated coding of patient discharge summaries using conceptual graphs. Methods Inf Med 1995 September; 34(4):345-51.

24. Schroder M. Knowledge based analysis of radiology reports using conceptual graphs. Proc the 7th Annual Workshop on Conceptual Graphs, Santa Cruz, USA; 1992. p. 445-70.

25. Chu S, Cesnik B. Knowledge representation and retrieval using conceptual graphs and free text document self-organization techniques. Int J Med Inf 2001 July;62:121-33.

26. Jiang MF, Tseng SS, Tsai CJ. Intelligent query agent for structural document databases. Expert Systems with Applications 1999;17:105-13.

Address of the author:
Stephen Chu, PhD, FACS,
Associate Professor of Health Informatics
Department of Management Science
and Information Systems
University of Auckland
Private Bag 92 019
Auckland, New Zealand
Email: stephen.chu@auckland.ac.nz