JPGM Journal of Postgraduate Medicine
Open access journal indexed with Index Medicus & ISI's SCI

E-MEDICINE
Year : 2000 | **Volume** : 46 | **Issue** : 2 | **Page** : 116-22

Information retrieval in medicine: overview and applications.

PM Nadkarni
Centre for Medical Informatics, Yale University School of Medicine, New Haven, Connecticut 06520-8009, USA. , USA

Correspondence Address:
P M Nadkarni
Centre for Medical Informatics, Yale University School of Medicine, New Haven, Connecticut 06520-8009, USA.
USA
Login to access the email ID

Source of Support: None, **Conflict of Interest:** None

PMID: 0011013482

Keywords: Databases, Bibliographic, Information Storage and Retrieval, Support, U.S. Gov't, P.H.S.,

How to cite this article:
Nadkarni P M. Information retrieval in medicine: overview and applications. J Postgrad Med 2000;46:116

How to cite this URL:
Nadkarni P M. Information retrieval in medicine: overview and applications. J Postgrad Med [serial online] 2000 [cited 2020 Dec 17];46:11
from: https://www.jpgmonline.com/text.asp?2000/46/2/116/298

:: Introduction and history

Information Retrieval (IR) is a branch of computer science that is concerned with the processing of collections of documents containing "fre
of hospital discharge summaries, radiology reports, or surgery notes, or (in a non-medical context) the full text of the complete works of Sha
spreadsheet or a database table, which is divided into rows and columns, such documents have no obvious structure: any structure that is imp
of little use in retrieving information. Thus, while the Bible is divided into books, chapters and, occasionally, verses, a reader is more interes
contains (e.g., the name of particular Biblical characters or a quote). Similarly, the "structure" of a Chest X-Ray report, in terms of the headi
different from the structure of a Barium swallow or Intravenous Pyelogram report.

IR began as an offshoot of "information theory," a field defined in a classic paper by Claude Shannon of Bell Laboratories in 1949. (Shanno
general audience.) However, "information" was defined in a very broad sense. Some of the work in this field considered practical problems s
Ziv and Lempel[2]) and how to add redundant (extra) information so that data transmission or storage would be reliable despite the presence
transmission, (e.g., Reed and Solomon[3]). The focus on textual information can be traced to several researchers, most notably the late Gera
the definitive textbook for IR[4]. Salton's group has performed research (using a system called SMART[5]) that has defined the methodolog
textual component of the clinical patient record, IR techniques have found ready applicability in medicine: a good reference for medical IR i

For several decades, IR was a relatively "orphan" technology researched by a relative handful of scientists, and commercial offerings were r
However, due to the spread of the World Wide Web, IR is now mainstreaming because most of the information on the Web is textual. Web s
and AskJeeves are used by millions of users to locate information on Web pages across the world on any topic. The use of search engines ha
access to the Internet, the World Wide Web has replaced the library as the reference tool of first choice. Search engines can also be purchase
search facilities for their own Web sites. For example, Microsoft bundles a component called Microsoft Index Server with its Web server sof
MEDLINE access as implemented through PubMed. (A subsequent article in this series will discuss MEDLINE article retrieval from the use
principles underlying IR technology.)

:: Document indexing

Search engines rely on pre-processing a document collection for subsequent fast retrieval based on keywords. This pre-processing step is ter
by the words contained in them are described below. Some of the operations involved are also employed in concept-identification algorithm

* Every document in a collection is scanned word-by-word, skipping words belonging to a stop-word list. *Stop-words* are very common wor
for searching (e.g., "the", "an", "of", etc). It is important to note that, in specific contexts, words uncommon in the language as a whole may
notes, "surgery" would be a stop word. Conversely, when processing medical text, certain common words should not be treated as stop word
concepts. Thus, "greater" and "lesser" specify a femoral trochanter.

* One or more disk-based structures (*indexes*) are created. The *global word-frequency index* records, for each distinct word, how many time
word-frequency index records how often a particular word occurs in each document. An optional *proximity index* records the position of ind
sentence, or paragraph offsets. (A proximity index takes up lots of space, but is useful for searches where one can specify two or more word
sentence, or next to each other, or within so many words of each other).

* The index stores words either as they occur in the text, or after transformation. One type of transformation is *normalization*, which involve
variations in person or tense. Thus, "children" is normalized to "child," and "brought" is normalized to "bring." Another transformation is *st*
suffix removal or letter substitutions to yield a "root" form. Stemming is more drastic than normalization; often yielding a root that is not a v
"hypothesize" and "hypothetical" yield the same root, "hypothes." Stemming uses simple linguistic rules (e.g., removing suffixes like "-ing"
on synonyms (e.g., liver, hepatic) is not used in stemming. The most widely used technique for stemming is the Porter algorithm[7]. Both no
but at the cost of specificity. (In the context of IR, *sensitivity*, also called *recall*, is defined as the ratio of the number of relevant documents i
query to the number of relevant documents *actually present* in that collection. *Specificity*, also called *precision* in the context of IR, is define
documents retrieved by a query to the *total* number of documents retrieved: some retrieved documents may turn out to be irrelevant to the qu
tests in diagnostic medicine, the higher the sensitivity, the less the *false negative* rate: similarly, specificity is negatively correlated with the *f*
is known to have problems in stemming words with Greco-Latin suffixes (-um, -ae, -ii, -us, etc.), which often occur in medicine. The C lang
described in[8] is driven by rules that are stored in tables, however, and such suffixes can be managed by adding more rules.

* For a large document collection, it is wasteful of disk space to store the words themselves in the index. Instead, a table is created to store e
word is assigned a numeric ID. This ID is used in the index.

* Every document in the collection is also given an ID. In the medical field, each document applies to a particular patient, and so one mainta
This information is used as a bridge between the free text and other non-free-text information (e.g., laboratory test values) on that patient, w
record system.

:: Types of query methods

There are two methods of searching documents using such indexes. The conceptually simpler (and older) *Boolean* method (named after the
Boole, who devised a mathematical approach to logic) allows the user to specify keywords combined with the operators AND, OR and NOT
the subsequent article on MEDLINE.) Boolean methods have been (falsely) criticized because Boolean logic is hard for end-users to underst
criticism of a purely Boolean approach is that the returned documents are not ranked in any way: if numerous documents are retrieved, the u
how relevant it is to the query.

Modern IR search engines (e.g., Web-based engines), which may retrieve hundreds of matches for a user's query, therefore perform *automat*
handle situations where no document matches every single keyword specified by the user. (In the case of Boolean search, zero documents w
still return partially matching documents.)

Relevance ranking is typically done through the *Document Vector approach*, originally devised by Salton's group. Several variants of this ap
between the terms in a query, and a document that must be screened to see how closely it matches the query, is called the *"normalized cosin*
metric is computed will not be described, the principle is that keywords in the query that are uncommon in the document collection are give
common, and that documents that contain one or more of the keywords in the query multiple times are given more weight than documents th
all. Thus, if we take a query with the keywords "carotid endarterectomy complications" the first two words would be less common, and ther
Similarly, documents that contained the words "carotid" and "endarterectomy" many times would be likely to have these words as a theme o
relevant to the query than documents that mentioned the phrase in passing only once.

The word "normalized" in the metric refers to the fact that the frequency with which a particular term occurs in a document is adjusted for th
abstract that mentioned "carotid" three times would be more relevant than a 10,000-word article that mentioned it four times. The normalize
documents for similarity: PubMed does this, as described later. The word "cosine" in the metric implies the metric is a number varying betw
two documents are identical with respect to the terms they contain and the frequencies of those terms, the "angle" between them is zero (and
in common (i.e., they are completely dissimilar), the "angle" between them is 90 degrees (and cos(90)=0). While relevance ranking sounds v
programming turns out to be quite simple, and relevance ranking means sorting the matching documents by descending order of cosine.

Modern search engines use a combination of Boolean and vector methods, with the former being used as a filter to restrict the search. Most s
word (with no intervening space) to mandate that all retrieved documents must contain that word. In other words, + implies the AND operat
indicates that the retrieved documents should NOT contain that word. (PubMed uses a slightly different syntax, as a subsequent article will i
Boolean criteria are then ranked by relevance.

:: Automated concept indexing

Thus far, we have discussed electronic indexing of documents based on the *words* within them. For documents belonging to a single domain
capitalize on domain knowledge. For example, synonymous phrases are not automatically recognized. Searches of medical free text that is ir
manually specify synonymous forms, or else risk missing relevant documents.

Using concepts in a domain-specific vocabulary (thesaurus) can enhance retrieval: one can index the *concepts* identified in a document. Inde
indexing by word. Text is processed to recognize potential terms, which are then looked up in a vocabulary. The ID of the concept in the voc
discussed in a subsequent article, the US National Library of Medicine has historically used manual concept recognition using trained humar
before assigning concepts to it. Here, we discuss *automated* approaches that have attempted to use computer programs for the same task.

:: Use of vocabularies in automated concept indexing

In medicine, concept-indexing research has used a variety of vocabularies, depending on the material that must be indexed. The biggest thesa
of Medicine's Unified Medical Language System (UMLS) Metathesaurus[11]. This incorporates numerous other vocabularies, such as SNO
used for pathological classification of disease), ICD-10 (Internal Classification of diseases, typically used for billing code), DSM-IV (Diagno
Disorders, 4th Edition), as well as MeSH (Medical Subject Headings), used by human MEDLINE indexers.

The UMLS's contents centre on concepts, terms, strings, and words. A *concept*, such as hypertension, may be expressed in different ways: e.
disease" and "hypertensive vascular disease" are different synonymous forms, or *terms*, that refer to the same concept. (Every concept is giv
expressed in multiple *string* forms through variations such as transposition of words, punctuation, or differences in case, person and tense, e.
is composed of one or more *words*. The UMLS contains numerous cross-reference tables that greatly ease the programmer's task. For examp
containing a particular word. A concept can belong to one or more *semantic categories* (e.g., pharmacological substance, therapeutic procedu
the *ID of the source vocabulary* (e.g., ICD-10) from which it was taken. This information allows researchers to create UMLS data subsets fo

:: Term recognition in automated concept indexing

We now discuss the methods used for automated concept indexing.

Phrase-Based Methods use natural language processing (NLP) methods to scan the text and identify words/phrases of interest, which are the
has utilized phrases representing nouns (e.g., "obstructive lung disease"), as in the work of Elkin et al[12]. Aronson and Rindflesch at the Na
approach, as described in several published papers[13],[14],[15]. If the entire phrase does not have a match in the vocabulary, one generates
vocabulary for matches against these subsets. Thus, "digitalis-induced atrial fibrillation" does not match any concept in the UMLS, but the s
do.

A problem with using noun phrases alone is that, in medical text, many concepts can only be correctly identified through other parts of speec
pressure was *greatly elevated*", which implies hypertension, as opposed to blood pressure alone. Verb phrases such as "surgically resected" a
excluded. Further, the same concept may be divided across two noun phrases, e.g., "hypertension is secondary to renal disease," which indic

In an attempt to overcome the limitations of single noun phrases, alternative approaches have attempted to use larger units of text, such as en
algorithms devised by Hersh[16],[17],[18]. Using larger text units yields more sensitivity, but at the cost of specificity; that is, there is an inc
text segment "spleen rupture and normal stomach" in an emergency surgery note will match the concept of "stomach rupture" which, while i
note. In addition, using an entire sentence is no guarantee of success. In the admittedly artificial example, "*Blood pressure* was recorded in th
elevated," the concept of hypertension is split across two sentences. Finally, the time required to generate subsets of the input phrase (and sea
the number of words in the phrase, and can be prohibitive for large sentences. Production systems therefore tend to favour phrase-based appr

:: Practical limitations of automatic concept detection in medical text

For medical text, detection of a concept in a document does not, per se, make that document relevant for that concept. The concept may refer
ruled out, or that occurred in the remote past. The recording of "significant negatives" is important in medicine, and robust recognition of ne
challenge.

Concept recognition addresses the problem of synonyms (different phrases for the same concept) but fails in the presence of *homonyms* (wor
example, the term "anaesthesia" by itself can indicate either a procedure for pain relief, or a clinical finding of loss of sensation. *Disambigua*
from several possible matches) can be performed based on neighbouring words. Thus, the phrase "endotracheal anaesthesia" can be uniquely
disambiguation is present further away, e.g. in the previous or succeeding sentence, the problem is much more difficult. Elaborate rules have
words that would favour one concept over the other. For example, mention of an anatomic location (such as the dorsum of the foot) would fa
clinical finding. However, the approach of manual rule creation does not scale up: the 1999 edition of the UMLS lists 13,688 ambiguous tern
manually would be a Herculean task. There may be other methods, however, that are less labour intensive than manually devising rules (e.g.

The presence of elisions, neologisms, or abbreviations in medical text complicates automated concept recognition. An *elision* is a phrase wit disambiguate the phrase based on the phrase's context. For example, in the phrase "the white count was 1800", "white count" is an elided fo a newly coined word that would not be expected to be in a dictionary. An example is "coumadinize" for the act of administering coumadin (c attained. Abbreviations that are also words in their own right can confuse a computer program, e.g., "RAT" can mean "recurrent acute tonsil numerous entries for elisions and abbreviations, but is by no means complete: it is therefore updated continually, with new data being releas

Some of the practical problems faced in automated concept indexing are discussed in[19].

:: Ir and traditional databases

While medical text is part of the patient record, it must coexist with more structured parts of the record that are stored in tables of a tradition demographic information or laboratory results. Historically, however, such integration has been lacking, and hospital-based software develo Over the last two years, the situation has improved considerably as relational database system vendors have integrated IR technology into th version 7 incorporates a full-text indexing engine, as does the "ConText option" of Oracle versions 7 and greater. Such integration enables th management.

Feature-wise, however, these offerings currently lag greatly behind those of companies that specialize in offering IR technology alone. For e provides the search engine for the well-known Physician's Desk Reference (PDR) database, goes beyond Microsoft and Oracle in several wa thesaurus, as well as allowing proximity searching (which has been discussed earlier in the section on document indexing). While it is possib engine if programming expertise is available (several engines, such as INQUERY[20], are freely available for researchers), it is best to go w purposes. The commercial engines have various extra features that would be highly tedious to program by oneself. For example, they unders HTML, MS-Word, and PDF). Other packages, such as IBM's Intelligent Miner for Text[21], constitute a full-fledged development environm a "Web crawler" that will search the Web on a periodic basis for keywords of interest, and a relational database engine where the results that

For many queries of medical data, it makes sense to combine a query of data present in traditional relational tables with that present in free t patients of a particular age, sex, and diagnosis who presented with particular clinical findings. Age, sex, and diagnosis are stored in tradition may be stored in narrative. Because querying relational tables for laboratory test values or diagnosis codes is much more efficient than query be used to reduce the number of documents that must be searched to identify patients matching a particular set of criteria. Vasanthakumar et query of traditional and textual data[22]. MS SQL Server and Oracle allow integrated query through proprietary extensions to the database s

When a medical institution is computerizing its electronic records, one must have a means of accessing old paper-based data that may go ba information into a computer by hand is extremely labour-intensive, and *document management systems* (DMSs) allow one to scan the paper referenced against essential demographic data on the patient such as medical record number, date of admission/discharge/death, primary dia to hand-enter this demographic data.) If the information on paper has been typed or printed rather than hand-written, one can often use *optic* images. OCR generates text files from the images, and one can then use IR technology, which is bundled into several DMSs, to index these f

:: Other areas of ir applicability

We have already discussed the document vector approach for relevance ranking. It is similarly possible, using the technique discussed earlie as one compares a document vector with a query vector. When one gathers a series of documents (passages) written by a particular author, th some extent collectively acts as a literary fingerprint for the author, because particular authors tend to use some words more often than other the proportion of verbs with a passive voice, can also be used for fingerprinting.) This knowledge has been used in various ways.

If the field of literature, if one collects several passages from different authors, document-vector methods can be used to *cluster the documen them. With enough passages per author, one typically observes documents from a particular author falling into their own special cluster. Thi suitably long passage by an author whose identity was uncertain (but about whom it was known that s/he was one of the profiled authors) do reasonable guess as to the author's identity. This technique has been used in detecting literary forgeries (as in the "Hitler diaries", where th in his autobiography, "Mein Kampf"). Similarly, a comparison of passages of William Shakespeare with those of Francis Bacon indicates th Shakespeare (as has been claimed by some). During the Cold War, "Kremlin Watchers" in the CIA were able to classify a particular commun one of a small handful of anonymous writers based upon which cluster the communiqué matched most closely. (These writers' identities wer names.)*

The Pubmed database, maintained by US National Centre for Biotechnology Information (part of the National Library of Medicine), provide Web. It uses an algorithm devised by NCBI's John Wilbur to compute similarity. For every article in medicine, supercomputer-class machine MEDLINE are similar to it (in terms of passing a statistical significance threshold). This information is stored in the PubMed database so th one can instantly find out all other references in the database that have been determined to be similar, by clicking a button on the Web form. full text of the article where available electronically: otherwise, only the abstract is used. The similarity algorithm is quite powerful, and it h originally intended by Wilbur. The journal Science reported that a researcher using PubMed discovered a case of serial plagiarism[23]. Sev to papers published by a particular Polish scientist had in fact been translated into Polish without Acknowledgment, and passed off as the sc to identify similarities despite the fact that it was working with translations in the latter case.

:: *Conclusions*

Information retrieval technology is the hidden workhorse that underlies the powerful Web-searching facilities that all of us take for granted armamentarium of the developer of computer software. Just as relational database packages have become progressively easier to use and ac microcomputer database packages will include IR components.

:: *Grant support*
: NIH Grants R01 LM06843-01 from the US National Library of Medicine and U01 CA78266-03 from the US National Cancer Institute.

:: *References*

1. *Shannon C, Weaver W. Mathematical Theory of Communication. University of Illinois Press; Chicago, IL: 1963.*
2. *Ziv J, Lempel A. A Universal Algorithm for Sequential Data Compression. IEEE Transactions on Information Theory 1977; IT-23:*

3. *Cipra B. The Ubiquitous Reed-Solomon Codes. SIAM (Society for Industrial and Applied Mathematics) News 1993; 26.*
4. *Salton G. Automatic Text Processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley; Rea*
5. *Salton G. The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall; Englewood Cliffs, NJ: 19*
6. *Hersh WR. Information Retrieval: A Health Care Perspective. Springer-Verlag; New York, NY: 1996.*
7. *Porter MF. An algorithm for suffix stripping. Program 1980; 14:130-137.*
8. *Baeza-Yates R, Frakes WB. Information Retrieval: Data Structures and Algorithms. Prentice-Hall; Englewood Cliffs, NJ: 1993.*
9. *Salton G, Wu H, Yu CT. Measurement of Term Importance in Automatic Indexing. J Am Soc Inf Sci 1981; 32:175-186.*
10. *Wilbur WJ, Yang Y. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. Com*
11. *Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Methods of Information in Medicine 1993; 32.*
12. *Elkin PL, Cimino JJ, Lowe HJ, Aronow DB, Payne TH, Pincett PS, et al. Mapping to MESH: the art of trapping MESH equivalenc the Annual Symposium on Computer Applications in Medical Care; 1988; pp185-190.*
13. *Aronson A, Rindflesch T, Browne A. Exploiting a large thesaurus for information retrieval. In: Proceedings of the RIAO; 1994; pp*
14. *Aronson AR, Rindflesch TC. Query expansion using the UMLS Metathesaurus. In: Proceedings of the AMIA Annual Fall Symposiu*
15. *Rindflesch TC, Aronson AR. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. In: Proceedings - the Annu Medical Care.; 1994; pp240-244.*
16. *Hersh W, Greenes R. SAPHIRE--an information retrieval system featuring concept matching, automatic indexing, probabilistic retr Biomed Res 1990; 23:410-425.*
17. *Hersh W, Hickam D, Leone T. Words, concepts, or both: optimal indexing units for automated information retrieval. In: Proceeding Applications in Medical Care; 1992; pp644-648.*
18. *Hersh W, Hickam D. A comparison of retrieval effectiveness for three methods of indexing medical literature. Am J Med Sci 1992;*
19. *Nadkarni PM, Chen RS, Brandt CA. UMLS Concept Indexing for Production Databases: A Feasibility Study. J Am Med Inform As*
20. *Callan J, Croft W, Harding S. The INQUERY Retrieval System. In: Proceedings of the International Conference on Database and I 356.*
21. *IBM Corporation. Intelligent Miner for Text. 2.3ed; 1999.*
22. *Vasanthakumar SR, Callan JP, Croft WB. Integrating INQUERY with an RDBMS to support text retrieval. Bulletin of the IEEE Tec 19: 24-33.*
23. *Marshall E. Medline searches turn up cases of suspected plagiarism (News). Science 1998; 279:473-474.*

This article has been cited by

1 ***Negation recognition in medical narrative reports***
Rokach L, Romano R, Maimon O
INFORMATION RETRIEVAL. 2008; 11(6): 499-538
[Pubmed]

2 ***Matching clinical and biological needs with emerging imaging technologies***
Vannier MW, Staab EV, Clarke LP
PROCEEDINGS OF THE IEEE. 2003; 91 (10): 1562-1573
[Pubmed]

Print this article | Email this Article to your friend

Previous arti