

A NATURAL LANGUAGE INTERFACE FOR AN INTELLIGENT DOCUMENT EVALUATION AND ANALYSIS SYSTEM

JAN G. WILMS UNIVERSITY OF MISSISSIPPI

Abstract -- The performance of a retrieval system can be enhanced by integrating conventional Information Retrieval methods with techniques adapted from Artificial Intelligence. A prototype of such an Intelligent Information Retrieval System has been built at the University of Mississippi, using the determinants of medication subject domain as an application area. This paper explores how a natural language interface can increase recall and precision by eliminating the need for a formal query language and the use of trained intermediaries to formulate a request.

Introduction

An Information Retrieval System (IRS) is a software package that analyzes a user's formatted query statement, and attempts to satisfy his information need by identifying a relevant subset of a document collection. Every IRS consists of an interface component that analyzes and interprets the user's query, and a retrieval component that lists references of documents that best match his request.

In the early sixties, H.P. Luhn introduced statistical approaches in the analysis and retrieval of documents to increase the efficiency and performance of retrieval systems [1]. Since then his technique has been perfected and implemented in commercial retrieval systems, but some fundamental issues still go unanswered. As one reviewer put it,

We [still] do not know the best way of representing the content of text documents and the users' information needs so that they can be compared and the relevant documents retrieved. We cannot even agree on a definition of relevance [2].

Besides this general feeling of dissatisfaction with the current state of affairs, there is a proliferation of computer systems that deal with text and on-line documents, which has led to an increasing awareness of the importance of IR as an application area.

Intelligent Information Retrieval Systems

As a result, there has been a recent shift away from traditional retrieval systems and their statistical approach toward something called "Intelligent Information Retrieval Systems" (IIRS). Sparck Jones defined such a system in 1983 as a user's probably ill-defined request to a set of relevant documents [3]. In other words, an IIRS is a system that carries out intelligent retrieval. Using stored knowledge about its documents and about the user and his need, an IIRS **infers** which documents will help the user satisfy his information needs.

This new concept of an intelligent system shifts much of the attention to the user interface. Most users are unable to specify exactly the information they need, since this involves describing the very thing they do not know. Consequently the computer-human interaction becomes very important in formulating a

query, and the need for unrestricted natural language queries becomes evident. A related feature is the dynamic utilization of user feedback in automatic search query reformulation.

A second characteristic of an IIRS is the emphasis on using an inferential process to link queries and users. Van Rijsbergen defines retrieval as a process based on logic, and views the matching function between query and documents as a plausible inference [4]; thus the retrieval process can be seen as an uncertain implication between a document collection D and a request R ($D \rightarrow R$). In order to do this the system must “know” about its task, and incorporate knowledge of the document collection, of the subject domain, and of the search topic.

Artificial Intelligence and Information Retrieval

The realization of the importance of the interface and of stored knowledge has led to the use of methods and techniques from Artificial Intelligence in IR. AI systems are systems that emulate human cognitive skills, and have traditionally been concerned with knowledge representation, declarative reasoning (in Expert Systems) and human interaction (through Natural Language Processing). The overlap between AI and IR has been approached from two angles: AI researchers using IR as an application area, and IR investigators blending traditional retrieval techniques with methods developed in AI research.

Expert Systems

One important area of AI that may benefit is expert systems. These are a type of knowledge based systems that are considered intelligent because they behave in such a way that a human behaving the same way in the same situation is considered to be intelligent. Thus

it may be possible to develop an expert intermediary system that assists the user with his query formulation, selection of search strategy, and retrieval evaluation.

I believe that expert systems research is the new frontier and the next area of development in library information science. Expert systems will enable users to make more effective use of the automated systems and on-line databases that were designed and implemented during the past decade, and they will help the libraries to be more productive and efficient in carrying out the many tasks involved in managing an information service center [5].

Expert systems work by observing human experts (like trained librarians) and deriving a set of rules and facts based on their expertise, which can guide the casual user and automatically refine his query. The system consists of a knowledge base and a set of facts and rules to traverse the search space.

Unfortunately, as several researchers point out, the success of this approach is disappointing, and the text processing context may not fit the proposed methodology; IR does not appear to be an ideal application domain for expert system development. In order to do any kind of intelligent problem solving of real-world tasks, expert systems require highly specialized domain knowledge, and hence are restricted to narrow specialist domain areas like diagnosing pulmonary disorders (PUFF) or configuring VAX 11 series computers (XCON). IR, however, is not narrow, nor homogeneous or well bounded. There are no obvious human experts, and there is no consensus on the best search technique. “Human intermediaries have no control over the retrieval algorithms employed by present systems and therefore treat them as a given, designing strategies to make use of their potentials and minimize their drawbacks” [6]. Search terms are often ambiguous and have

unclear relationships. Rules in information retrieval are not transparent and have consequences that do not follow unequivocally from the antecedents. Moreover, there is much evidence that the traversal of a hierarchically structured knowledge base shows little resemblance to the actual search strategy used by human experts.

Natural Language Processing

An important feature of an IIRS, and another major area of overlap between IR and AI, is a flexible and convenient interface which allows powerful interaction between user and retrieval mechanism.

With the advent of interactive computer terminals in the early 1970s, it was expected that the on-line revolution would involve the end-user directly in the automatic search and retrieval process. This hope was not realized, however [7]; many competing retrieval systems were developed (DIALOG, STAIRS, MEDLARS), each with different access mechanisms and control languages that are incompatible. The standard Boolean search mechanism used by these systems is confusing to the average user because he is untrained in

using the operators in their strict logical sense. He often believes that “X AND Y” will retrieve more documents than “X” alone and, forgets that “X OR Y” will rate documents with one term equal to documents containing both items.

Because of these user-hostile systems, casual users have to rely on trained search intermediaries to guide them in their query. The real query is in the user’s mind, and most of the understanding process in traditional systems happens off line through interaction with expert searchers. These intermediaries must assume the full burden of understanding the user’s information need, and coming up with the “right” search strategy and query formulation. “Existing IR systems are basically passive, ‘dumb’ systems searched by dynamic, ‘intelligent’ human searchers” [8].

Following is a sample of such a complex query, and how a trained searcher would convert it to a Boolean formulation: “Excretion of phosphate or pyrophosphate in the urine or the effect of parathyroid hormone on the kidney” -> SELECT (URINE AND (PHOSPHATE OR PYROPHOSPHATE)) [9].

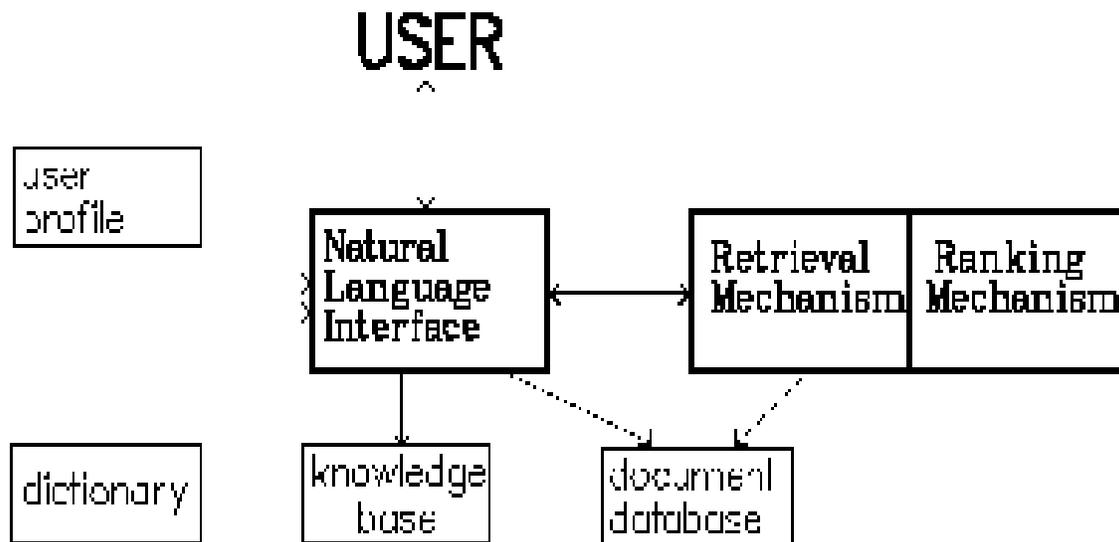


Figure 1. System Architecture

\$ID 590001
 \$TI Prescription Patronage Motivation
 \$AU Ohvall, Richard A.
 \$SO M.B.A., University of Wisconsin, 1959
 \$YR 1959
 \$CC 101.01; 209.02; 302.08; 402.02; 408.01; 409; 601;
 \$AB Personal interviews and structured questionnaires were used to gain insight into the factors that influence patients to have their prescriptions filled in one particular pharmacy rather than another. ...

Figure 2 Document Sample

Several operational systems have tried to overcome this complexity by using a simplified command language and yes/no menus and scripts. The emergence of user friendly interfaces and gateway systems include such automatic features as dialup and logon procedures, saving search statements, and assistance through help screens and on line tutorials. But all these fixes still offer very little assistance in the formulation of the query.

The heightened subject diversity and text volume present more vocabulary switching problems and create a need for more advanced NLP capabilities. To meet this need, new releases and versions of operational IR systems [...] exhibit even more powerful and versatile functions in the area of proximity searching, automatic multifile query transformation, and multilevel user interfaces, but they still shy away from formal language-analysis techniques and tools. [8]

The proposed approach for an IIRS is to allow the user to enter his query using natural language, and to build an interface that will map his input to a traditional query language. Many researchers agree to the superiority of using “high-level query languages which reflect the overall user’s intent rather than the computer operations that may be required to obtain any particular result” [10].

Natural language processing (NLP) can play a role in both the retrieval and storage of documents. It can be used to build a friendly user interface that allows free language query submission and hence eliminates the need for mastering a formal query format. For infor-

Field	Value	Weight	Threshold
\$ID	590001	(1.0)	(1.0)
\$AU	Ohvall, R. A.	(1.0)	(1.0)
\$YR	1959	(1.0)	(1.0)
\$TI	Prescription Patronage Motivation	(1.0)	(1.0)
\$CC	101.01; 209.02; 302.08; 402.02; 408.01; 409; 601	(1.0)	(1.0)
\$AB	Personal interviews and structured questionnaires were used to gain insight into the factors that influence patients to have their prescriptions filled in one particular pharmacy rather than another. ...	(1.0)	(1.0)

Figure 3 Query Template

mation storage, it can be used to structure the document database and perform automatic indexing. In addition, NLP can help construct synonym dictionaries and thesauri.

Some researchers have attempted the automatic formulation of Boolean queries from natural language without the use of linguistic input or grammatical theory [9]. Instead of semantic criteria, they use probabilistic tests; the algorithm translates a request by using the estimated number of documents retrievable by a term or term combination until a predetermined threshold has been reached. The estimation is based on the sum of posted document frequencies of individual terms or terms combinations. The output has been tested on the MEDLARS system and is clearly competitive with conventional manual Boolean formulations, and even superior when good natural language statements of user need are available.

NLP consists of several levels of conventional processing, some more relevant to IIRS than others. The phonological level analyzes speech sounds, and is of minor importance in IR, except maybe for generating sound alike words in approximate name matching (e.g. if the author’s name is misspelled). The morphological layer performs operations on word stems, and can be used to remove suffixes, allow truncation operations, and permit browsing through alphabetically arranged

entries (e.g. ELLHILL's "neighbor" command). Lexical operations include full word processing, and are useful in stopword deletion, spell checking, automatic search key substitution or augmentation, or handling of abbreviations and acronyms through thesauri. The next higher level up is syntactic identification of structural units like noun and verb phrases. Sophisticated automatic parsers have been developed, but are rarely used in operational systems, except occasionally in free text searching (adjacency and pattern matching) and search restriction to certain boundaries (e.g. to Title or Author field). Finally there is the semantic category which uses contextual knowledge to represent meaning. Here no formal methods exist and there is currently little use for it in existing IRS. (The nearest approximation to it would be the display of classification schemes like ELHILL's "tree" and "explode" commands). (There is one higher "pragmatic" level in NLP which uses knowledge about real-life objects to make meaning unambiguous, but this is very experimental at best even in AI research).

The use of linguistic methods to analyze natural language input of user queries has both opponents and proponents. Some individuals believe that the analysis of meaning improves the retrieval process, and consider information retrieval as an early stage of more refined question answering [11]. Others doubt the usefulness because of the difference between information retrieval and other areas of language processing. Like the critics of the expert system approach to IR, they value much more the use of statistical, probabilistic or vector space techniques [12].

The whole idea of meaning representation is dubiously relevant to document retrieval in anything like its present form. [...] One can get quite good results with simple terms and weights [13].

This study sides with the researchers who believe that well established linguistic

procedures do contribute to retrieval effectiveness. It is true that unrestricted natural language input still poses formidable problems because of inherent semantic ambiguity and absence of a general theory of speech acts and dialog pragmatics. However, an interface based on augmented transition network parsing and fuzzy set techniques is proposed, which is able to handle constrained natural language queries and thus eliminates the need for mastering a formal query syntax.

Intelligent Document Evaluation and Analysis System

At the University of Mississippi, a project has been undertaken in cooperation with the department of Health Care Administration to develop an IIRS. The system uses the "determinants of medication" as an application field, which is a well defined small interdisciplinary area dealing with the manufacturing, marketing, and consumption of medication. A document collection is being assembled with over three thousand items that are related to the broader domain of pharmacy and medicine. The goal of the system is to provide easy access to this document collection, and to research how IR can be enhanced by AI. The system is to simulate the expertise of a human information specialist who is trained in the manipulation of documents, without being an expert in the subject matter (determinants of medication). Thus the purpose of the project is to develop an operational IRS, and an experimental software testbed for research. Consequently the system is built around a very modular and flexible design.

Architecture

The system consists of 2 integrated components (interface and processor), each embodying some AI techniques to enhance the recall and precision of the retrieval process. The

system architecture, which is based on a similar study by Biswas et al. is depicted in figure 1 [14]. This paper will focus on the first component, the natural language interface.

The architecture of the proposed IIRS reflects a knowledge-based system approach: there is a clear separation between the domain specific knowledge base (topical area of medication) and the inference mechanism which does the retrieval and ranking.

Each document in the document data base consists of a template with 7 fields, as shown in figure 2.

The knowledge base, with its domain specific terms and relations, is completely separated from the inference mechanism, so the IIRS can easily be used with new domains. It is partitioned into three components: a concept knowledge base (for fields \$AB and \$TI), a chronology base (for \$YR), and a name base (for \$AU and \$SO). Each knowledge base entry consists of a list of terms that represent meaningful entities and a set of operators to combine these concepts; the concept knowledge base, for example, consists of a thesaurus of terms related with synonym operators (e.g. “antacids” = “adsorbents”) and through the implication relation (e.g. “aminoglycosides” -> “antibiotics” -> “anti-infective agents”). This thesaurus has been carefully built by an expert in the field of medication, and will be used in helping to reformulating the query if it is unsuccessful. It also allows much more flexibility in the original formulation of the query since the user can use familiar terms, and the system will enhance the query automatically (non obtrusively), as a specialized librarian would do. Similarly, the chronology base contains synonyms (“about” and “around”), and establishes concrete values for fuzzy specifications (e.g. “recent” = 1980 +). The name base, finally, resolves abbrevia-

tion and acronym ambiguities, and deals with misspelled names through approximate name matching.

The natural language interface uses a mirror image of the template to analyze the user query. This image of the user’s information need can be conceived as an ideal hypothetical document, perfectly relevant, or as an idealized representation of a set of items the user wants to retrieve. The input is partitioned into the different fields (\$AB, \$AU, \$SO, \$TI, \$YR). After the initial query, the system negotiates with the user to fill in the missing fields. This interaction continues until a threshold is reached that indicates enough data is available, or until all fields are filled. The \$ID field will remain blank, of course; this is how the inference mechanism returns a ranked list of document numbers.

The IIRS keeps a user profile which retains information concerning the user’s interaction with the system. Through profiles an IIRS can demonstrate machine learning and be more sensitive to the user’s specific information need and manner of query formulation. Profiles delimit the portion of the document space normally searched in respond to a query, and influence the query negotiation process. The use of profiles is especially viable in systems with relatively fixed groups of users, with stable individual interests, but breadth of interest within each group [15]. Since the IIRS is developed for a document collection of interdisciplinary determinants of drug use (sociological, medical, economical), it is not uneconomical to maintain a separate database for each user. Expert users, for example, will probably have a relatively complete initial query, so the query negotiation to fill blank fields will rarely be necessary. Hence the profile can be used to dynamically set the template threshold. User topology (expert or beginner) can usually be inferred from the

queries themselves (e.g. generality or specificity of concepts).

The inference mechanism uses a standard inverted file and Boolean operators to retrieve a set of relevant documents, and ranks them by means of weighted request and document processing. This ranked list of documents and their weights is returned to the user via the \$ID field of the template (See Figure 3). It is the responsibility of the language interface to display the references to the user. Hence the interface component can be considered the scheduler of the system. In addition to being ranked by weight, the documents are released to the users one by one, each time accompanied by a request for relevance feedback to enable on the fly computation of document similarity.

If the retrieved set of documents proves unsatisfactory, the IIRS will assist in reformulating the query. Each field has a weight which indicates how many documents were eligible for retrieval for this subset of the query. If the retrieved set of documents is too narrow, for example, and the \$AU field contains "1", then this is an indication that so few items were returned because the query was too specific about the author. The user can modify the previous request in a way similar to "query by example;" the frame structure is displayed and the user can modify each field and resubmit the query.

Fuzzy Logic and Augmented Transition Networks.

In stead of a crisp query language, the IIRS recognizes imprecise natural language queries. To develop a sufficiently robust and friendly interface, it is unavoidable that some ad hoc restriction in permissible input is imposed.

A pragmatic approach to designing NL processors would involve tailoring them to interpret statements and terms only in the context and domain under consideration. This reduces language ambiguities and makes the implementation task easier [14].

Only a single imperative sentence is accepted, but these represent the most common query form. Example: "Give me very recent survey articles written by M. Smith about both pediatric medication use and therapeutic substitution". The query consists of crisp items (pediatric, medication, therapeutic, Smith), imprecise terms (recent), and fuzzy qualifiers (very). The last 2 are considered fuzzy because they convey imprecise information, and do not have sharp distinctions between membership or non membership to classes. To handle these uncertainties, each concept is given a weight, which is determined by fuzzy logic [16]. Since the inference mechanism works only with OR and AND operators, the NOT operator is simulated by giving a negative weight. In accordance with the strict logical meaning, "and" is interpreted as OR unless it is modified by "all," "both," or "neither." In this case a logical AND is simulated by assigning high weights to the operands, so they will be ranked high in the retrieved set. Similarly, compound terms like "therapeutic substitution" are kept together by high individual weights.

The first step in the translation process is morphological; suffixes are removed from the input string, and word stems are generated using an on line dictionary. Next comes the lexical operation: a stop list is employed to delete common words (not including modifiers or conjunctions). This step also includes spell checking and thesaurus operations. Syntactic parsing follows; grammatical rules are used to build a structure that depicts the relations between words in the sentence. There are several established algorithms to achieve this:

context free phrase structure grammars (S -> NP + VP), context sensitive transformation grammars (surface structure mapped to deep structure) and Augmented Transition Networks (ATN) [17]. Rather than taking a deep-understanding approach, the natural language processor uses a pattern-matching like method for parsing (e.g. "Give me" corresponds to NIL, "about" introduces the \$AB field). The IIRS relies on users not trying to confuse the system with excessively ambiguous or complex queries. To insure correct interpretation the NLP unit always displays its analysis so the user can correct any mistakes before the query is processed. Thus the natural language interface infers content attributes and external attributes (e.g. \$YR) and places them in the correct field.

Conclusion

An intelligent information retrieval system provides enhanced performance by integrating conventional IR methods with techniques adapted from artificial intelligence. A natural language interface can virtually eliminate the need for a formal query-formatting syntax and the formulation of a request as logical document set manipulations of specific terms. By restricting the user's input to single imperative sentences, the interface can process the user's query statement without requiring infallible detection of syntactic errors or semantic ambiguities.

References

- [1] Luhn, H.P. "A Statistical Approach to mechanized encoding and searching of Literary Information." IBM Journal of Research and Development. 1 (1957) : 309-317.
- [2] Croft, W. Bruce. "Approaches to Intelligent Information Retrieval." Information Processing and Management. 23 (1987): 249-254.
- [3] Sparck, Jones K. "Intelligent Retrieval." Intelligent Information Retrieval: Informatics 7. Jones, K.P. (ed). London: Aslib, 1983: 136-142.
- [4] Van Rijsbergen, C.J. "A new Theoretical Framework for Information retrieval." Proceedings of the ACM Conference on Research and Development in Information Retrieval. Pisa. Italy (September 1986): 194-200.
- [5] Borko, H. "Getting Started in Expert Library Research." Information Processing and management. 23 (1987): 81-88.
- [6] Brooks, H.M. "Expert Systems and Intelligent Information Retrieval." Information Processing and Management. 23 (1987): 367-382.
- [7] Salton, G. "Expert Systems and Information retrieval." ACM SIGIR Conference Proceedings. 21 (Spring/Summer 1987): 3-10.
- [8] Doszkocs, Tamas E. "Natural Language Processing in Information Retrieval." Journal of the American Society for Information Science. 37 (1986): 191-196.
- [9] Salton, G.; Buckley, C. and Fox, E. "Automatic Query formulations in Information Retrieval." Journal of the American Society for Information Science. 34 (1983): 262-280.
- [10] Salton, G. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983.
- [11] Gardin, J. "On the Relation between Question Answering systems and Various Theoretical Approaches to the Analysis of Text." The Analysis of Meaning. M. MacCafferty and K. Gray (eds): Aslib: London, 1979: 206-220.

- [12] Robertson, S. "Between Aboutness and Meaning." The Analysis of Meaning. MacCafferty, M and Gray, K. (eds). Aslib: London, 1979: pp 202-205.
- [13] Sparck Jones, K. "Problems in the Representation of Meaning." The Analysis of Meaning. MacCafferty, M. and Gray, K. (eds) Aslib: London, 1979: pp 193-201.
- [14] Biswas G. et al. "Knowledge-Assisted Doc Retrieval: I. The Natural Language Interface." Journal of the American Society for Information Science. 38 (1987):83-96.
- [15] Korfhage, R. and Chavarria-Garza, H. "Retrieval Improvement by the Interaction of Queries and User Profiles." Department of Computer Science and Engineering. Dallas, TX: Southern Methodist University, 1982.
- [16] Zadeh, L. "PRUF--A Meaning Representation Language for Natural Languages." in Fuzzy Reasoning and its Applications. Mamdani, E. and Gaines, B. eds. New York: Academic, 1981.
- [17] Winnograd, T. Language as a Cognitive Process: Syntax. Reading, MA: Addison-Wesley, 1983.