Commission of the European Communities

information management

The application of recent software technology to the access to patent information systems

Report

EUR 11326 EN

Commission of the European Communities

information management

The application of recent software technology to the access to patent information systems

D. Vermeir, E. Laenens, J. Dierick

Dept. of Mathematics and Computer Science
University of Antwerp, UIA
Universiteitsplein 1, 2610 Wilrijk, Belgium

Directorate-General
Telecommunications, Information Industries and Innovation

1987

EUR 11326 EN

Published by the
COMMISSION OF THE EUROPEAN COMMUNITIES
Directorate-General
Telecommunications, Information Industries and Innovation
Bâtiment Jean Monnet
LUXEMBOURG

LEGAL NOTICE

Neither the Commission of the European Communities nor any person acting on behalf of the Commission is responsible for the use which might be made of the following information

Cataloguing data can be found at the end of this publication

Luxembourg: Office for Official Publications of the European Communities, 1988

ISBN 92-825-7773-2 Catalogue number: CD-NA-11326-EN-C

Printed in Belgium

CONTENTS

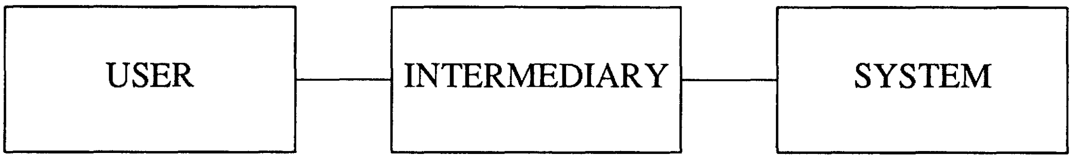
1. INTRODUCTION

The advent of online information retrieval systems (OIRS) has made it possible for the interested user to have instant access to a vast amount of 'knowledge' which is stored in a growing number of public databases.

Unfortunately, this important resource is underused for a variety of reasons, the most important one being the necessity of mastering a number of technical skills required to make effective use of the different systems.

Among the problems a user of OIRS faces, we mention the selection of (and connection to) the appropriate database, gaining familiarity with the intricacies of the information representation on a particular system, the design of a good search strategy and the formulation of the search in the formal language of the system.

Not surprisingly then, end-users are generally unwilling to search personally but instead rely on the assistance of an information retrieval (IR) expert. This professional, who is often not an expert in the subject matter of the search, acts as an intermediary between the user and the system, as is illustrated below.



Clearly, this state of affairs is far from ideal:

(i) Many potential users who could benefit from OIRS cannot afford to employ or consult such an intermediary.

(ii) If the intermediary is not a domain specialist as well, chances are that important information will be overlooked.

Among the various OIRS, **patent information systems** (PIS) are probably among the most underused: although they contain a wealth of knowledge which is useful for many purposes, they are structured in a way that makes it especially difficult for a novice to carry out a meaningful interaction (see below).

Recent developments in software technology, namely in the area of artificial intelligence (AI), have made it feasible to construct 'intelligent' systems that allow a non-technical user to perform tasks that normally require an expert.

Such so-called **expert systems** have been successfully used to assist with medical diagnosis, trouble shooting in a variety of environments, chemical analysis, network access and management etc.

In view of the above, it then becomes natural to try to employ expert systems technology to automate the role of the information retrieval intermediary.

In fact, there have been several proposals in this direction [1,2] (see Section 3 for an overview).

In this report we present a preliminary study of the feasibility of an automated expert assistant for accessing patent information systems.

Such a system would combine expertise in the following areas:

(i) which kind of information is available on which host and at what price;

(ii) how to connect to a particular host;

(iii) the formal language and dialogue model employed by each host;

(iv) the structure of patent information;

(v) the effectiveness and use of existing tools such as the international patent classification (IPC) codes, the catchword index, DARC etc.;

(vi) search strategies;

(vii) domain knowledge, including relevant terminology in several languages;

(viii) knowledge about the user profile (preferences etc.).

The main requirement of the proposed system is that a domain expert, e.g. a mechanical engineer, without prior information retrieval experience, should be able to access patent information systems, using the proposed system, as effectively as with the help of a human intermediary.

The remainder of the report is organized as follows: in Section 2, we discuss patent literature and existing access tools and methods. We also present user experiences with patent information systems. From these experiences we can draw certain conclusions for the requirements of the proposed expert system.

Section 3 contains an overview of recent developments and proposals in the field of information retrieval, with particular emphasis on the proposed applications of expert system technology in this area.

Section 4 gives a general description of the functionality and architecture of a proposed expert system for patent information retrieval.

Finally, Section 5 presents conclusions and recommendations for further research and development. The appendix contains an annotated bibliography.

2. PATENT INFORMATION SYSTEMS

2.1. Patents

A patent is in fact a contract between the applicant (the inventor or a company) and the government or authority: the patentee receives the monopoly (production, trading, licensing, etc.), over an invention protected by government through jurisdiction, during a certain period (between six and 20 years). In return, the patentee must publish his knowledge and has to use or let use the invention. Consequently, society benefits through this agreement. One should keep in mind the double meaning of the word 'patent': on the one hand it means the exclusive and temporary exploitation rights, guaranteed by the government. On the other hand it can be used for the paper document itself on which the right is written down. In the context of this study, we consider as 'patent document': all primary documents issued during the granting procedure (from application to granted patent, novelty search reports, etc.), reissue patents, defensive publications, etc., i.e. the official documents issued by the government or certain international organizations, in order to publish and register the rights whether they are granted or not. They contain a description of technical objects or methods and claims. The secondary patent literature is formed by the reference and abstract lists, used to disclose the primary documents.

Since its origins in Western Europe (fifteenth century) , the patent legislation is still developing and is adjusted to the changing technological and sociological environment worldwide. From 1987 onwards a completely new patent law will be in force in Belgium; there is a new Chinese patent law; the protection of computer software and chip-industry is under way and a European community patent will probably be installed in the near future. There is even a movement to create a uniform worldwide patent application and granting procedure.

2.1.1. Differences between patent and non-patent literature

Non-patent literature, which deals with the conventional periodicals, monographs, etc. is used far more frequently than patent documentation. Non-patent literature is not suitable for an exhaustive state-of-the-art search, because plus minus 70% [3] of the information described in patents is never published in periodicals or books. Only half of the remaining 30% is cited fully elsewhere.

An explanation for this phenomena is given by (Oppenheim, 1979):

(i) an inventor does not take time to write an article; he rather continues his research;

(ii) patent documents are sometimes quite voluminous: the patent BP 749836 holds 267 pages and 780 drawings.

One should not omit patent literature when following the development in a certain technological area. There is always a time gap between the writing of a document and the moment of its publishing: this gap is on average smaller with patents because an inventor wants to protect his research results as quickly as possible. Most authorities publish their patent documents within 18 months after application. It is remarkable, for instance, that the description of the computer punched card appeared in a patent document 25 years earlier than in conventional literature; for the jet-motor this delay was 10 years, for television 5 years.[4]

In the context of 'universal bibliographic control' patent literature could stand as an example: there is a worldwide standardization (formats, codes, exchanges, lay-out of documents, classification, identification, microforms, digitalization, etc.) which is still growing due to the efforts of the World Intellectual Property Organization (Geneva), and there are reference lists issued by the official authorities giving a 100% disclosure and a maximum reliability.

The yearly growth of published patent documents is about 1 million, i.e. the same number as for conventional literature. It is expected that there will be 17 million basic patents in 1995. The danger exists that this huge collection will crash under its own weight and will be totally uncontrollable. This could be avoided through intensive automatization and higher use of abstracts and indexes. The growth rate is exponentional. A sample, taken in US/CLAIMS, by means of the yearly number of chemical American patent documents in the period 1950 up to 1985, indicates that:

$$p(t) = 1301.66 * e^{\frac{t}{12.93}}$$

where

t = year (take e.g. 50 for 1950)

p(t) = number of patents in year t.

However, due to the recent economic recession, one should be careful with the extrapolation of this sort of statistic.

2.1.2. Contents of a patent [5]

Due to the international standardization almost all issued patent documents have nearly the same lay-out:

(a) The bibliographical part (on the front page):

(i) identification (country, numbers, etc.);

(ii) classifications (IPC and/or others);

(iii) kind of document (ICIREPAT-codes);

(iv) dates (priority, application, granting, etc.);

(v) title, abstract,

(vi) names (inventor, applicant, agent, etc.).

Each field has a unique identification code (ICIREPAT- INID-codes):

(b) A description part:

(i) state-of-the-art;

(ii) problem and solution;

(iii) explanation of the drawings.

(c) The claims (main and secondary claims).

(d) Drawings.

2.2. Importance of patent literature

Patent documentation is the best source for:

(i) technology transfer: a company benefits from its own R&D results through in-house use or through licensing, or can use the research results of others;

(ii) competitor monitoring;

(iii) avoiding 'reinventing the wheel': double work in R&D is estimated as 30%;[6]

(iv) trend analysis, which is becoming more and more indispensable in a company or country investment policy.

Metaphorically speaking: 'with a book or article one can bark; with a patent one bites'.

2.2.1. Reasons for the unpopularity of patent documentation

An American inquiry [7] states that only 4.8% of the interviewed engineers find patent literature very important. It is 'the ugly duckling' in the documentation world.

Often mentioned reasons are:

(i) patents are better known as legal documents rather than as an information source;

(ii) patent texts are often hard to understand immediately, due to the 'patentese': technical jargon, confusing terminology and long complex sentences, especially in the claims;

(iii) education in patent use, during the training of young technicians and managers is insufficient;

(iv) the availability of patent documents is lower because large patent libraries are costly;

(v) the instant usefulness of information contained in a patent is not always obvious. Ready-to-use solutions for technical problems are seldom given: a patentee wants protection as much as possible but at the same time hides his research results for his competitors.

2.3. International patent classification (IPC)

The purpose and meaning of the IPC could be resumed as follows:

(i) a worldwide standard classification;

(ii) a system to classify and arrange the storage of a huge collection of documents so that all documents concerning a subject are close together (especially for novelty searches);

(iii) it facilitates the disclosure of patents in foreign languages;

(iv) it can be used for OIR (novelty search, selective dissemination of information, state of the art, etc.);

(v) it is suited for trend-analysis.

Experienced IPC-professionals state that the accurate use of this classification is not easy. Some of them, even after years of intensive use, frequently confront discussion and problems when assigning an IPC-code to an invention. Consequently, it is easy to understand that the novice IPC-user will need close assistance, and that he must take the time to browse and evaluate the possibilities.

The number of patent issuing agencies giving IPC-codes to their patent documents is constantly growing while the use of similar national classification systems is diminishing. In addition, the CAPRI-project (i.e. an INPADOC-WIPO-cooperation) is reclassifying patent documents according to the IPC down to 1920. This recoding will be completed in 1988, so that all important patents (PCT - minimum collection) will be retrievable via an IPC-symbol (12 million documents).

This classification scheme is specially conceived for and by patent experts. 'The classification, being a means for obtaining an internationally uniform classification of patent documents, has as its primary purpose the establishment of an effective search tool for the retrieval of patent documents by Patent Offices and other users, in order to establish the novelty and evaluate the inventive step (including the assessment of technical advance and useful results or utility) of patent applications.' [8]

IPC-codes can be used in online information retrieval.

IPC is established in the English and French languages, but translations in other important languages exist (German, Russian, Spanish, Japanese, etc.).

The whole body of knowledge is divided in 9 sections, and these are broken down into classes, subclasses, groups and subgroups. This elaboration results in over 58.000 subdivisions.

A complete classification symbol comprises the combined symbols representing the section ('A'), class ('01'), subclass ('B') and maingroup ('1/00') or subgroup ('1/24'), e.g.

A 01 B 1/00
or
A 01 B 1/24

The hierarchy among groups is determined by the number of dots preceding the titles of the subgroups. These dots are also used in place of, and to avoid repetition of, the titles of hierarchically directly superior groups:

example: A 63 H 3/00 Dolls
3/36 . Details; appurtenances
3/38 . . Dolls' eyes
3/40 . . . movable

Operational rules (use of references, notes, standard expressions, etc.) are very important for the accurate interpretation and use of the fields.

The principles of the classification, i.e. the difference between 'function-oriented' (the intrinsic nature or function of a thing), and the 'application-oriented' (the particular use or application) approach, and the classification of different aspects of an invention (process, compositions, details, etc.) demand an expertise or a thorough assistance for proper selection of a classification code, both for indexing and retrieval.

An X-notation (i.e. an 'X' added to an IPC classification symbol) is used when the subject, mostly dealing with new boundaries of technology, cannot be classified satisfactorally under the present elaborations. The retrieval of such X-subjects, using IPC is very difficult: a match between the indexers and retrievers use of the X-symbol (and above all which hierarchical level is appropriate) is hard to expect.

Since the fourth edition of the IPC there is a possibility to add to the obligatory classification (dealing with the essential aspects of the invention), non-obligatory classification or indexing codes, which describe additional useful information. These supplementary codes, forming the so called 'hybrid system' improve retrieval features.

Every five years there is a revision of the IPC-elaborations. Revision-concordance-lists (IPC2/IPC3 and IPC3/IPC4)show which IPC-codes are revised and what has changed. For an exhaustive search one has to examine the patent literature using previous IPC-editions! Concordance-lists between the IPC and several national classification systems (German, American, Russian, etc.) are available.

An official catchword index facilitates the entry to the classification scheme.

How to use the IPC as a search tool? The IPC can be used in an infringement search, a novelty search, a validity search, an informative search, etc.

One can use the following procedure:

(i) describe the relevant field(s);

(ii) identify the proper place(s) in the classification; use the catchword index by means of specific nouns (process, device, product, etc.) or consult the 'content of section' and follow and select the elaborations;

(iii) all IPC-rules have to be borne in mind, and all additional remarks in the classification have to be incorporated;

(iv) select the relevant subclass(es);

(v) check scrutinously the notes hereunder and execute the instructions;

(vi) locate the appropriate main group using the subclass index;

(vii) scan all one-dot groups and select the most relevant one;

(viii) the group-code to be used in the search is the one which is most indented (most dots) but which is still covering the subject. Follow all the notes and instructions;

(ix) consider using the hierarchically higher groups (up to the main group level);

(x) evaluate the found documents and if necessary repeat the IPC examination.

Some examples of IPC-rules which can be automated

(a) As mentioned previously, when reading a subgroup-title, one has to follow the dot-indexed indentations.

Example:

A 01 B 1/00 Hand tools
1/24 . for treating meadows or
lawns

The title of 1/24 is to read as: 'hand tools for treating meadows and lawns'.

A 01 B 1/00 Hand tools
1/16 . Tools for uprooting
weeds

The title of 1/16 is a complete expression, but owing to its hierarchical position, 'the tools for uprooting weeds are restricted to hand tools'.

(b) Interpunction in the titles (separation of 2 or more distinct parts separated by semicolons).

Example:

B 64 Luftfahrt; Flugwesen; Raumfahrt

(c) Scope of places at any level, e.g. the scope of a main group is to be interpreted only within the effective scope of its subclass.

(d) Instructions and notes following the titles.

(e) Last place rule.

(f) References: three different functions are possible:

(i) limitation;

(ii) precedence note;

(iii) guidance.

(g) Special wordings; i.e. , e.g. , per se , covered by, covered in, or the like, etc.

(h) The hybrid system: obligatory/non-obligatory codes and their separation by special interpunction.

(i) How to define the balance of invention when there are two or more equal main aspects; there is a fixed order prescribed.

The advantage of IPC

The biggest advantage of the use of an IPC-code as a key when performing a subject search, is that c.q. the 'recall' (theoretically) reaches 100%, because all indexers assign the same IPC symbol to the same invention. If errors occur when assigning the original codes, correct retrieval is of course not possible; but let us assume that the indexers are competent.

Opposite to the exclusive use of IPC codes there is the exclusive use of keywords: in that case, the recall will not even approach the maximum, because for one subject there are always many keywords and combinations of them relevant: it is very unlikely that the searcher will use them all. Consequently, one loses a part of the recall.

Some problems related to keyword search are:

(a) what to search?
  - (i) the original title,
  - (ii) the enriched title,
  - (iii) the abstract,
  - (iv) the full text,
  - (v) the additional keywords,
  - (vi) or a combination of the above;

(b) which keywords to apply?
  - (i) free words,
  - (ii) controlled thesaurus words;

(c) different spellings of:
  - (i) synonyms,
  - (ii) 'strange words',

(iii) proper nouns.

The 'noise' has an inverted behavior: when using IPC, it is rather high because the elaborations are not detailed enough for narrow questions. When using keywords, noise could be low, although the used keyword may be assigned by indexers to irrelevant contexts, which pushes the noise up again. The solution lies, as often mentioned by the interviewed practitioners in the inquiry, with the combined use of IPC and keywords. One could propose the following method:

(i) select via an IPC symbol;

(ii) eliminate the family members and withhold only the 'basics' (if needed);

(iii) when there are still too many hits, combine very carefully with keywords, each complemented with all its synonyms (possibly after use of the ZOOM test, or consultation of the thesaurus).

The user could search the appropriate IPC-symbols using different methods:

(i) starting with a keyword, going via the catchword index, which leads directly to the relevant elaboration;

or

(ii) starting in the classification at the highest level ('classes') and going through all the hierarchical levels, each time selecting the relevant underlaying elaboration.

The first method has the advantage of speed, because the number of choices is limited and the searcher goes directly to the right place, where maybe some nuances must be pointed out. The disadvantage here is the selection of the right entry: the 67 000 entries of the catchword index will often prove to be insufficient. The proposed structured thesaurus could reduce keyword selection difficulties. The second method offers educational qualities because the user runs through the basics of the IPC, and has to apply the IPC philosophy at each level and choice. Here there are no problems with selecting keywords: the title and notes of each elaboration explain the underlaying content. The disadvantage here is the relatively long search time (five to six levels to run through) and the danger of going totally wrong after an early mistake at a higher level.

2.4. User experiences

This was investigated by means of a limited inquiry.

2.4.1. Purpose of the inquiry

The purpose was to determine, through a dialogue with a number of online patent searchers, the present methods, problems and possible solutions concerning online information retrieval in patent files.

A brief description of the functionality of the proposed expert system was given and the interviewed practitioner was asked for a general appreciation of the proposed interface.

2.4.2. Method

In August 1986, five large Belgian companies were asked to cooperate with the inquiry. All five reacted positively. An appointment was made by telephone with the company's online patent searcher. An explanatory letter, accompanied by the questions to be asked, was sent to them for personal preparation. All five interviews, conducted in a half structured way, backed by the prepared questions, were positive and satisfactory. Afterwards, an acknowledgement and a copy of the final report will be sent to the cooperators. Among the visited companies, two are chemical, two mechanical and one in the photographic industry. Four of them are quite large multinationals with an intensive R&D activity, with their own inhouse patent agent(s) and with a well-equipped documentation centre where the competent online searchers function.

2.4.3. General remarks

All of the visited companies keep their own patent data base (in paper and/or computerized forms) in order to have rapid and easy access to the information. One firm even cooperates with its competitors in establishing their private database (though selectively accessible via Dialog, Lockheed). They use deep indexing for disclosure. As source they use the official national bulletins and online patent files; relevant hardcopies are ordered and stored.

Online information retrieval is believed to be very expensive, when all the factors are kept in mind:

(i) equipment (soft- and hardware);

(ii) PTT-costs (and especially in Belgium);

(iii) host-subscription and search costs;

(iv) and above all: human cost, not only for the search time, but for the time needed to keep pace with the ever changing landscape of the online business.

The yearly online budget in one company runs into millions of Belgian francs.

Patent information, especially when used for juridical purposes, has to be 100% accurate. So online answers have to have a maximum recall. The rate of precision is in this context less important.

Although all interviewed searchers are experienced (several searches per day), it was obvious that they all had different skills and accents. This could be altered by the proposed interface when the expertise is kept central.

The patent information is 'polluted' by the Japanese deluge (500 000 patent documents per year published).

Bibliographic searches (names, dates, numbers, etc.) are rather simple to formulate and to retrieve. Subject searches are the hardest to cope with: one has to find as much as possible relevant keywords and/or classification codes, and then relate them in a boolean combination.

There are some difficult subject areas, like chemistry, mechanics and electronics because the most relevant information is contained in a drawing. Intelligent retrieval interfaces like 'DARC' (Telesysteme-Questel) and 'TOPFRAG' (Derwent) are very popular among the experts.

Most online searches are centralized in the visited firms. One person conducts the search mostly in (spoken or written) dialogue with the person who wants the information.

In the company context, not only the patent agents ask for (online) patent information. The commercial, marketing, management, documentation, etc. divisions are also users of patent information.

2.4.4. Used hardware

Personal computers, with printer and in one case direct communication to the central inhouse database.

Two searchers used a communication software package for:

(i) offline search preparation;
(ii) choice of host and file;
(iii) automatic connection;
(iv) automatic downloading, etc.

2.4.5. Files and hosts

2.4.5.1. General remarks:

(i) Little use is made of files beyond the classic patent databases like the Derwent files, US/CLAIMS, INPI 1-2-3-4, and the INPADOC files, etc. The choice is a function of the desired information. One has to be skilled to make the right selection.

(ii) All desired information is online available but dispersed. In this context crossfile searching is wanted, (this is already achievable with ESA-IRS: WPI and CAS, though only for bibliographical data).

(iii) The 'ZOOM'-function on ESA-IRS is a useful tool.

2.4.5.2. Some detailed remarks concerning specific producers

Derwent files: (World patent index and world patent index latest)

(i) suitable and cheapest for family searches (references to equivalents are given);

(ii) good patent oriented abstracts;

(iii) enriched titles (improves the readability and understanding at the stage of browsing);

(iv) better user feedback via Derwent agents in several countries and regular user meetings;

(v) the coverage is not complete (28 countries);

(vi) for chemical searches: fragmentation codes-system is error-inviting and complex. The TOPFRAG - interface will solve this inconvenience;

(vii) there is a useful home-made thesaurus.

INPADOC:

(i) much too few data: only the numbers and titles are given;

(ii) there is no family reference system: too many unwanted equivalents;

(iii) the coverage is far better: 50 countries;

(iv) expensive;

(v) only subject retrieval through IPC: the precision is too low.

2.4.6. Problems

The most frequently formulated problems concerning online patent searching were:

(a) the different hosts who all create:

  (i) different command languages,

  (ii) different interpretations of searchable fields and their pre- or suffixes (e.g. author - assignee);

(b) insufficient cross-file-search-possibilities;

(c) different or wrong spelling of names (firms, persons, etc.);

(d) the high cost fo an online connection (hosts and communication);

(e) frequent updates:
  - (i) command languages,
  - (ii) file and record structure,
  - (iii) host-realia: address, new or disappearing, etc.;

(f) PTT-malfunctioning:
  - (i) message errors (idle or missing characters, etc.),
  - (ii) interruptions,
  - (iii) congestion;

(g) finding the exact search formulation in subject searches.

2.4.7. About the international patent classification

The overall opinion about the application of IPC-codes in subject searching, was that it is not very useful. The reasons are multiple: it lies with the IPC, with the indexer and with the user himself. The biggest drawbacks are:

1. The subdivisions are not refined enough, so there are too many references attached to each IPC-code (subgroups). The 'recall' could be quite acceptable, but there is too much 'noise' (i.e. low precision). In this aspect, the more refined EPO-inhouse version of the IPC (30% more elaborations), was thought to be more satisfactory, and its online release was eagerly expected. This refinement problem is most critical in the chemical field. The over 7 million existing substances can not be searched with an acceptable precision by means of IPC- subgroups. Only the input of the unique (sub-)structure identification (e.g. the Markush-formula) can help here.
2. Concerning the use of the IPC:
   - (a) some patent issuing agencies attribute IPC-codes to their patent documents only up to the 'main group' level;
   - (b) some documents receive too many IPC-codes describing their content (up to 35 codes), consequently retrieving exactly those documents only using their IPC-codes could be quite complex;
   - (c) in the indexing process:
     - (i) different indexers interpret differently the IPC- subdivisions despite its universal character,

(ii) completely false codes are often attributed so that in the retrieval process any matching is impossible.

3. The IPC in its full extent and with its revisions, is not easy to comprehend thoroughly, even for the experienced users. So searchers are reluctant to use it. Finding the most relevant classification code for a given subject is frequently rather complex: there could be disagreement whether the topic should be classified 'function-oriented' or 'application-oriented', the grade of refinement (main group or subgroup), etc. Some interviewed searchers advised not to search always down to the lowest classification level because the upper level could also contain relevant references.

The most frequently heard positive remark was that the IPC can be of help, when used in combination with keywords. The IPC is regarded as a 'first direction', which has to be augmented by descriptors for narrowing the subject field, or as the ultimate solution when all other retrieval methods failed (no hits or too many).

An ingenious way of bypassing the problem of finding the exact code is to use the codes written on a previously and otherwise found relevant document (whether or not checked before further use).

2.4.8. Expectations concerning the proposed expert system

As mentioned previously, two communication packages were used in the visited firms (e.g. 'Crosstalk'). The people who use them are satisfied with the given capabilities but are aware of its limitations: the nucleus of the search-act, the query-formulation, remains fully the searchers job. There is no assistance in this part.

The following functions were thought to be desirable for the proposed system:

(i) automatic logon and logoff procedures;

(ii) offline query formulation;

(iii) downloading and reformatting;

(iv) one simple dialog-language (e.g. menu-driven);

(v) deal with updates of command languages, file and host structure changes (is believed to be very difficult to achieve);

(vi) the expert assistant must run on a personal computer;

(vii) the expert system must use as much as possible the tools given by the hosts and files in order to perform a narrow search profile;

(viii) the expert system must use existing tools like DARC, TOPFRAG, ZOOM, PATSTAT, etc. which are too powerful and useful to omit or neglect;

(ix) the implementation of a built-in exhaustive thesaurus is not possible: e.g. in the chemical field: too many substances, synonyms, jargon names. Its updating would cause even greater trouble.

Some general considerations concerning the briefly described expert system: Is it really necessary that even the technical directors secretary is given the ability to search online for patent information. Can it be left to the classic human intermediary? The same remark sustains for the lab-researcher or the bench-engineer: why do they have to go online when in many cases there is an expert information officer in his neighbourhood?

Many information seekers do not really know exactly what they are looking for; the narrowing borders of the wanted documentation come only after an iterative process during the search- retrieve- evaluate- process. How can the expert system ever intercept this?

How far must the help offered by the expert system go? Technically one could take the searcher by the hand, but is it worth the effort?

As said before, online searching is an expensive activity, especially due to the time needed for consistent follow-up of the updates. An automatically updated assistant is very welcome, even if it is expensive.

2.5. Specific features of patent literature versus the classic documentation sources

(i) It is limited to technology.

(ii) The relevant online files are limited.

(iii) The bibliographic control is 100% (no Grey literature or Invisible Colleges): every published patent document worldwide is registered, numbered and available.

(iv) The presence of 'doubles' is mastered by the family system.

(v) In many cases patent retrieval demands 100% recall (juridical disputes).

(vi) Patentese.

(vii) Drawings are absolutely necessary.

(viii) Patent files must be modifiable: patent data are often altered (legal status, family members, etc.).

3. INFORMATION RETRIEVAL

3.1. Information retrieval - general

3.1.1. Databases - databanks

Due to the present knowledge explosion - this term rightly describes the uncontrollable character of the massive literature production - the digitalization process became a necessity. One tries to master this deluge of documents by putting their references in the 'secondary' literature: lists of pointers to the original publications, equipped with several entries and more or less complete abstracts. Full text capturing is the main objective for the future. The computer files, produced by specialized publishers, containing these references, are called 'databases'. At this moment there are about 2 764 databases (Levy, 1986). There is a fundamental difference between 'fact-finding-databases', and the above mentioned 'reference-databases'.

Host computers make (a selection of) the databases available: these are the 'databanks'. Now there are about 414 of them (Levy, 1986).

3.1.2. Communication

The link between the computer at the users' site and the host computer runs via the common PTT-network and a special network. The messages are electronically transmitted, e.g. by a modem. The searcher, having a password and an identification number, can search the files in the host computer when he pays for it. These costs differ in function of hosts, file, question, etc., and can increase easily (often due to PTT-cost, especially in Europe). The searcher must converse with the host computer in a rigid command language. These differ strongly from host to host. The European 'common command language' tries to solve this difficulty.

3.1.3. Information retrieval systems

Information retrieval systems are designed to search files of information and retrieve stored documents or references to documents in response to queries specified by a user. Typically, the stored items are described by words contained in the document texts, sometimes supplemented by additional related information. These words, index terms, are called controlled terms if they are assigned by professional indexers and free terms if they are derived automatically from titles and abstracts of documents. Queries describe a user's need for information and often consist of index terms interrelated by Boolean operators. The retrieval system acts as a filter, selecting documents whose characteristics match the query specification.[9, 10]

3.1.4. Automatic indexing

Automatic indexing strategies make possible the design of effective automatic-text-based retrieval systems that are fully competitive with conventional manual operations and can be operated without the need for human subject or domain experts for document indexing and search formulation.

A basic procedure for automatic indexing could be:

(i) Identify the individual words occurring either in the documents or in document excerpts (e.g. titles and abstracts).

(ii) Use a stop list of common function words (and, of, or, the, etc.) to delete from the texts the high-frequency function words that are insufficiently specific for content representation.

(iii) Use a suffix stripping routine to reduce the remaining words to word stem form; this recall-enhancing transformation broadens the scope of the terms and can be performed automatically using a limited number of basic rules.

(iv) For each remaining word stem i occurring in document j, compute a term weighting factor, which is the product of the term frequency of term i in document j multiplied by the inverse document frequency of term i in the collection as a whole. Available evaluation results indicate that term weighting improves retrieval effectiveness by distinguishing the important content terms from the less important ones.

(v) Represent each document by the chosen set of weighted word stems.

Of course, this basic indexing process can be improved by adding some refinements.

3.1.5. Effectiveness measures

The effectiveness of a retrieval system is usually evaluated in terms of a pair of measures, known as recall and precision. Recall is the proportion of relevant material actually retrieved from a file, while precision is the proportion of the retrieved material that is found to be relevant to the user's needs.

In principle, a search should achieve high recall by retrieving almost everything that is relevant, while at the same time maintaining high precision by rejecting a large proportion of extraneous items.

In practice, it is known that recall and precision tend to vary inversely, and that it is difficult to retrieve everything that is wanted while also rejecting everything that is unwanted. A very specific query formulation produces high precision and hence low recall performance. As the query formulation is broadened, more relevant items are retrieved, thus improving the recall, but also more non-relevant ones, thereby decreasing the precision.

When a choice must be made between recall and precision, most users choose precision-oriented searches where only relatively few items are retrieved and the user is spared the effort of examining a large amount of possibly irrelevant material - the penalty attached to a high recall search.

In automatic retrieval systems, both query formulation and document representations can be altered to reach the desired recall and precision levels through the use of recall-enhancing devices (e.g. term truncation) to broaden the document and query identifiers, and precision-enhancing devices (e.g. term weighting) to make item identifications more specific.

3.1.6. Search techniques

A search technique uses the information derived in the indexing process to compute a similarity value or a probability for each document, specifying how likely it is to be relevant to the query.

Term weights enhance the search precision by distinguishing the better, or more important, terms from the less important ones. Such a discrimination may also help rank the output in decreasing order of presumed importance.

The use of ranked document output improves the user-system interaction by alerting the user to the more important documents first. Information culled from the documents retrieved early in the search can then be used to generate improved query formulations in subsequent searches. This process is called relevance feedback.

Some of the search strategies used in current systems are:

(i) Boolean search in which the query formulation is constructed using logical combination of query terms;

(ii) cluster hypothesis which postulates that the documents relevant to one request are on the whole more alike, than they are like non relevant documents;

(iii) interactive search formulation in which the requester becomes involved in a trial-and-error process in which he reformulates his query on the basis of an answer to the initial query;

(iv) probabilistic retrieval approach in which the number of times a term or concept appears in a document is a significant feature of relevance.

3.1.7. User interface

Some of the assumptions made in the design of current commercial systems are that the user has a definite idea of what he is looking for, and that he has expressed this need precisely in the vocabulary of index terms allowed by the system. This is usually not the case. Either the user has only a general idea of what he wants, cannot properly reduce his need to appropriate index terms, or both. In any case, the results produced by the system can be less than adequate.

The available systems respond to each query and each user in the same way. No provision is made to use more than just a superficial indication of the user's information need. Furthermore, these systems have no way to incorporate any background information, which is relevant to the search but not directly to the need.

To get the most complete results possible, the user may have to use several systems, which means he has to learn the intricacies of several command languages.

Intermediaries

The usual solution to the problem of getting effective results with commercial retrieval systems is to employ people called search intermediaries. Generally, intermediaries are not experts in the subject area being searched. They are experts at using the retrieval systems. The intermediary's task can be divided into a number of steps.

(i) Get from the user a characterization of his information need, either by interviewing him or by having him write it down.

(ii) Select the appropriate database.

(iii) Reduce the need to a query appropriate to the database being searched.

(iv) Use the system to find the documents.

(v) Present the documents to the user for his evaluation.

(vi) Go back to the first or second steps, if necessary, to revise the approach, and repeat until the user is satisfied.

The intermediary must be both a negotiator and a strategist. Negotiation describes the interaction between the user and the intermediary during the interview. The strategist aspect describes the intermediary's interaction with the retrieval system.

However, there are disadvantages with using the services of a search intermediary. The main one is that the person with the information need is not using the system. The intermediary, who does the actual search, has only a limited knowledge of the domain being searched. It is impossible for the searcher, who has the knowledge of the domain, to describe all relevant information to the intermediary.

Besides, one of the obvious criteria of information retrieval systems is that they must be directly accessible by the end user. There is, unfortunately, no place for a human intermediary in every sitting room.

User friendly systems

On the other hand, human beings can be very good at helping one another to solve problems. So a potential approach to the problem is to design systems which simulate human intermediaries in some ways, i.e. to make the system more flexible and responsive such that the need for a human intermediary can be removed. Flexibility can be incorporated into a number of aspects of these systems such as:

(i) Query formulation: by giving the user more assistance as he begins to formulate a query, and by letting him browse the document collection, he can develop a more precise query, thereby retrieving relevant information more quickly.

(ii) Retrieval techniques: there is no retrieval technique which is the best for all queries. If a system could select the best retrieval technique for a specific query, the performance of the system would be significantly improved.

(iii) Interface design.

3.1.8. Future of OIR

The online business is developing quickly. Due to the new and cheap storage techniques, which must be accompanied by new and powerful retrieval-tools (e.g. the proposed system), the documentation world will completely change. Especially for patent documents one sees the development in the 'legal- status-information', the growth of mixed files (patent and other information), the storage of full text documents combined with graphics (see PATDPA, in the FR Germany) and the improvement of chemical substructure-searching.[11]

3.2. Problems and user friendly criteria

The command language used in online information retrieval systems is typical of the available methods of communication between the human and the computer. The instructions given are completely unambiguous; either the command is in the correct format and is obeyed, or the system does not recognize the command and control is returned to the user with an error message. Such a dialogue reflects a number of communication problems:

(i) many details must be mastered by the user before the system becomes an effective tool;

(ii) the system is 'brittle', unable to function with even minor variations in human input;

(iii) communicating with the system has little continuity or sense of direction;

(iv) the user is not helped to decide what to do next or how to get the system to achieve particular purposes.

While existing systems are quite similar in their functional capabilities, the similarity across systems comes not from a common understanding of human expression but from a common approach to storing data and an interest in bibliographic citations. Retrieval systems have proved unsatisfactory for a number of reasons:

(i) the relationship between a researcher and the body of relevant printed knowledge was not understood;

(ii) there are very few cases where a complete and exhaustive bibliographic search is necessary, yet most devices are designed on this basis;

(iii) the intellectual access to most systems is so artificial and difficult that they invite neglect.

As seen in the inquiry conclusion and confirmed by the literature study, some more concrete constraints are:

(a) the great number of hosts, each with their own command language which is frequently updated and refined. Nobody can learn all the needed details by heart and one has to maintain a vast amount of system documentation;

(b) there are too few cross-file-search capabilities; i.e. the transfer of search results in one file to use them in another file. Although some hosts do have this capability;

(c) the different spelling of nouns;

(d) OIR is expensive;

(e) modifications in the online landscape:

  (i) new hosts or files,

  (ii) retro-active updating ('backloging') of the files,

  (iii) changes in file and record structures, addresses, codes, access protocols, etc.

(f) PTT-malfunctioning;

(g) the selection and combination of search terms (having the recall/precision ratio in mind);

(h) the right application of Boolean operators in complex searches;

(i) a novice searcher has to run through a long and difficult learning process with trial and errors;

(j) the strict grammar and terminology of the command languages and the harsh machine-human dialogue;

(k) the systems themselves do not help very much when there is a crisis situation;

(l) some typing skills are needed because an average search demands more then 100 strikes;[12]

(m) The person who wants information often does not know himself what he needs, or cannot find it in the right search terms;

(n) human intermediaries are often bad translators of the searcher's needs (different conception of the problem);

(o) when even a good search strategy produces too much noise, this is often the fault of the producers, because of wrong or insufficient indexing of certain aspects; e.g. in chemistry: the so called role indicators like pressure, temperature, function, etc.[13]

The proposed user friendly system will have to suppress as much as possible of the above mentioned difficulties or replace them with an adequate by-pass. E.g.

via detailed and readable and understandable assistance (adjusted to the experience of the user), via limited online costs by preparing the search off- line, etc.

In the field of AI, a lot of research effort has been dedicated to intelligent interfaces being able to communicate with a human. A graceful interaction is not a single monolithic skill but a number of diverse abilities.[14] Some criteria for user-oriented online information retrieval systems are:

(a) flexible parsing: the ability to deal with natural language with all the ellipses, idioms, grammatical errors and fragmentary utterances it can contain;

(b) robust communication: the set of strategies needed to ensure that a listener receives a speaker's utterance and interprets it correctly;

(c) focus mechanisms: the ability to keep track of what the conversation is about;

(d) explanation facility: the ability of the systems to explain what it can and cannot do, what it has done, what it is trying to do, and why, both for responses to direct questions and as a fall back when communication breaks down;

(e) identification from description: the ability to recognize an object from a description, including the ability to pursue a clarifying dialogue if the original description is unclear;

(f) the learning mechanism: the ability of the system to acquire facts, new skills and more abstract concepts from experience, and the ability to learn from its own mistakes;

(g) knowledge of the user: the system should have the ability to diagnose the level of the user and create his model;

(h) correction of errors: the users have to be provided with the opportunity to correct their errors or the system needs to be able to judge the validity of the input;

(i) user friendliness: the system should be user-friendly and easy to use;

(j) tutorial aids for users: the system should be equipped with tutorial modules;

(k) response time: the response time of the system should be adequate and the variance in response time should be minimized;

(l) search strategy: the intelligent interface should assist the user in search strategies. This includes being capable of being searched in ways that are unconventional but convenient for the user, e.g. once the user has located some relevant items he should be able to instruct the system to find others 'like them';

(m) ranked document output: be capable of weighting search terms automatically and ranking documents by degree of match with the search statement.

3.3. The relevance of AI technology for IR

Practical reasons for looking at AI in the context of computer-based information systems include the possibility of making systems accessible to a wider range of people, delegating certain tasks to the system while helping the user with more complex tasks.

3.3.1. AI applications

Four AI concepts have particular significance for information systems: pattern recognition, problem solving, representation and learning.

Pattern recognition

Pattern recognition is the identification of an object with a particular set of features as the member of some class. The problem of reference retrieval is similar to that of pattern recognition. In reference retrieval the development of document surrogates and query formulations may be viewed as a classification problem. The term 'features' is a useful generic term in the context of reference retrieval as well as AI, for it permits one to think of index terms, authors, citations, etc. all as possible features. Feature selection can occur in retrieval systems at two points: when document surrogates are prepared and when queries are formulated for comparison with document surrogates. Items in the file are classified in response to each query, when the portion to be retrieved is separated from that which is not. Although sorting as an approach to classification requires that the user specifies exactly which features must occur for an item to be retrieved, the query to find all items 'like' one already known to the user is a problem of prototype matching. The system assesses the probable relevance of a document to a query by calculating a measure of similarity between a document surrogate and the query formulation. An item is retrieved if the similarity measure is above some threshold.

Representation

The representation is a formalism for the knowledge possessed by a system. It may be thought of as 'a set of conventions about how to describe things'. Just as representation in AI is a formalism for knowledge possessed by a system, a document representation is 'a formalized statement of the nature of a document'. A query formulation may be viewed similarly. When one thinks of computer-based systems rather than manual, one must ask how to take the available information and represent it in a way that the computer can store and manipulate. This includes not only representations for documents and queries, but also relations between documents (such as citation relations) and between terms (such as those shown by a thesaurus). Online systems must be designed to encompass not only internal representations, i.e. representations of information within the computer subsystem, but also external representations - the displays of information at the user-computer interface.

Problem solving

Problem solving is the art of using knowledge effectively to attain desired goals. It can be approached using either algorithms or heuristics. In reference or data retrieval, the problem confronting the system is to identify, in response to each query, the portion of the contents of the file which should be retrieved. In this case problem solving includes development of a search strategy and the use of some inference mechanism. Application of heuristics could be the use of techniques which allow one to quickly select the subset of the file satisfying the query. Online systems must be designed to include consideration of how best to build the user-computer interface so that poorly constructed queries can be converted to well-structured forms that the computer subsystem can handle.

Learning

System elements such as feature selection routines, representations, and heuristics are of course all initially selected and programmed by a human designer when an AI system is developed for some application. Learning mechanisms by which a system can improve its performance over time are therefore necessary so that the initial design does not circumscribe system capabilities. The availability of online computer systems makes it reasonable to speak about dynamic systems which change and improve performance over time. Learning in retrieval systems can have either short term or long term effects. Short term learning is the modification of system response during the processing of a particular query in order to better meet the needs of the user. In reference retrieval systems, for example, this can be done through feedback in query processing, taking account of the relevance status of a sample of retrieved documents as judged by the user. Long term learning could involve modifying and/or extending the representation to improve system response over time. Modification can include changes in file organization and in item representation, e.g., updating the database to reflect new terminology. Extension can include techniques for storing previous search strategies in a form suitable for subsequent use by other system users.

What is particularly interesting about such knowledge based systems is that there are three different modes of end user behavior in contrast to the single mode for conventional information retrieval systems:

(i) user as a client: get answers to problems;

(ii) user as a tutor: improve the systems knowledge;

(iii) user as a pupil: harvest the knowledge base.

3.3.2. Expert systems

Characteristics of an expert system

There is often a difference between the technical definition of an expert system and the expectations of those who wish to use it.[15] People think of an expert system as one which will help them carry out tasks outside their own range of capability. But the technical definition is more concerned with the structure of the program which is being presented to the user. The features of an expert system are outlined here in a general way and then related to the problem of providing access to information systems for inexperienced users.

An expert system has the following components:

(i) a knowledge base which contains important information in the subject area and the connections between the different pieces of information;

(ii) an inference mechanism which is able to use the connections between the information to make conclusions or formulate advice to present to the user;

(iii) an explanation system which tells the user why certain actions were taken;

(iv) a system for adapting the inference mechanism to the experience gained from recording the user activity.

Automated searching as an expert system

As mentioned earlier, an expert system for information retrieval is required to mimic the capabilities of a skilled information professional. This involves many different facets:

(i) choosing a database which contains the required information;

(ii) choosing retrieval terms which describe the search topic;

(iii) knowledge of the retrieval language commands;

(iv) knowledge of the communication systems and protocols;

(v) ability to interact with the host computer dialogue;

(vi) ability to react to error messages;

(vii) ability to modify the search in the light of the results obtained from the information retrieval system.

The system proposed by Croft and Thompson is a nice illustration of an expert system for information retrieval. [9] The system helps a user formulate a query that specifies his information need and retrieves documents to meet this need from the available databases. We will look at this expert assistant in a next section. Another interesting expert system is the Userlink system,[15] also discussed later.

3.4. Recent developments in IR

3.4.1. An expert assistant for document retrieval

Croft and Thompson present the design of an expert assistant.[9]

The system provides information and tools to help a user formulate a query that specifies his information need, and provides a number of search techniques for retrieving documents to meet that need.

At the outset of his search, the user may have only a general idea of the information he wants, and may be unsure of how to specify it.

The system assists the user by gaining knowledge about him and his need, and uses this knowledge to guide the presentation of information for query refinement. While this interaction occurs, both the user and the system refine and expand what they know. The user gets a better idea of what his query should be; the system gets a better idea of both the user and the need.

When the information need has been made sufficiently clear, the system will select the most effective search technique to retrieve documents for the user evaluation.

The expert assistant is more than a conventional retrieval system: it can recognize when the user needs help and offer it, although he is not obliged to accept it. Furthermore, it can offer explanations of its actions at a level appropriate to the user.

The system is divided into three major components: the interface manager, the system experts and the knowledge base. The system experts are a collection of function specific experts which give the system the capability to assist the user in clarifying and expressing his information need and to retrieve documents likely to be relevant to that need. The knowledge base is the repository of all information collected by the system. The interface manager contains the knowledge needed to display and collect information in an appropriate form for a given device and a given user.

The system experts view a blackboard, called the short term memory, in the knowledge base and perform various actions depending on the current context. The experts also use information from the other part of the knowledge base, the long term memory. These two parts of the knowledge base represent what information applies to the specific session in progress and that which the system knows in general.

There are seven experts in the system:

(i) browsing expert;

(ii) explainer;

(iii) thesaurus expert;

(iv) request model builder;

(v) user model builder;

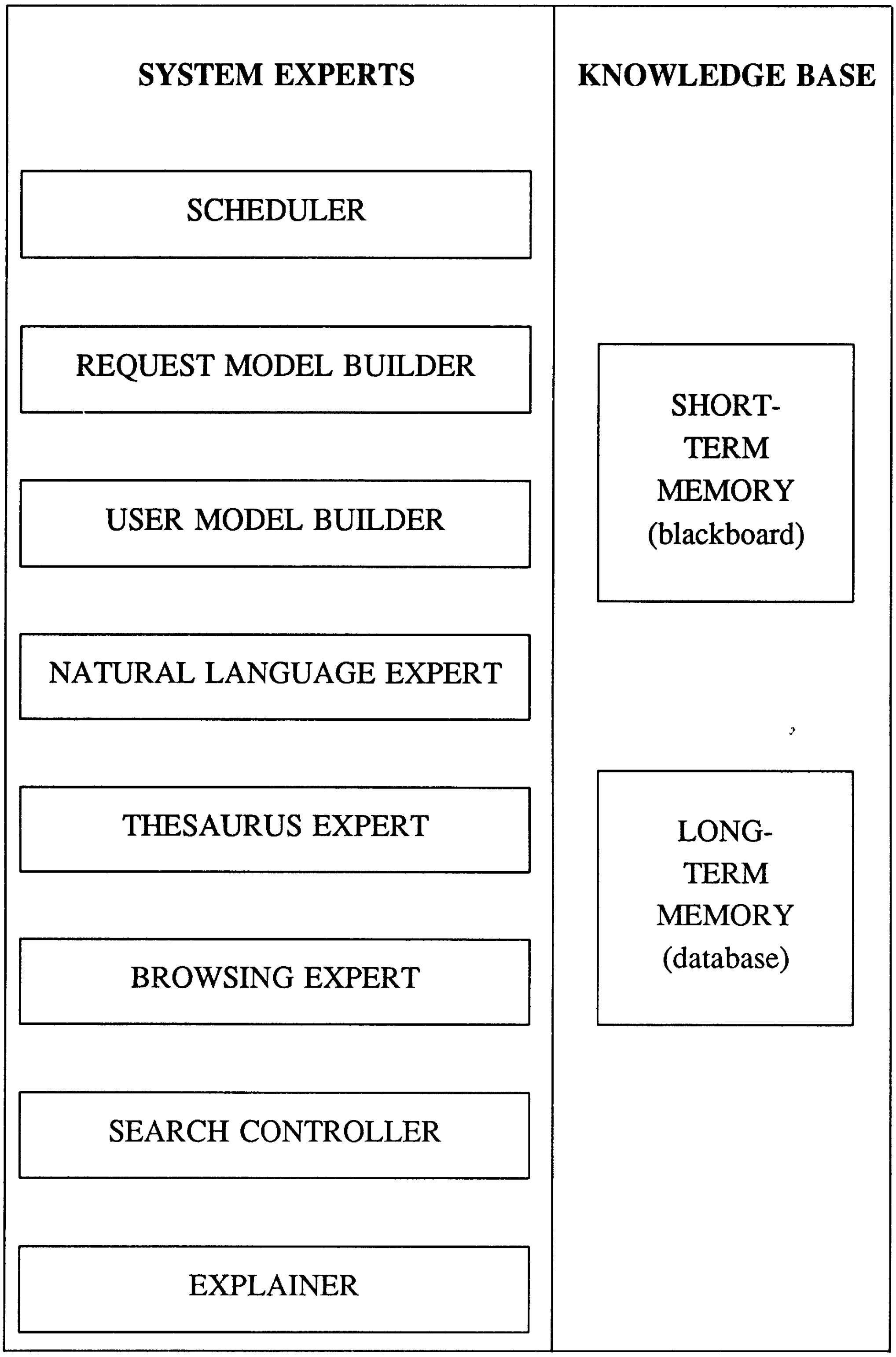
(vi) search controller;

(vii) natural language expert.

Basically, they fall into two groups: those with which the system initiates actions (3-7) and those with which the user initiates actions (1-2). The first group operates by asking the user for information, which they then use to build or modify models. The second group allows the user to interrogate the system for specific information about its knowledge base and its operation.

The scheduler is responsible for coordinating the activities of the system experts. In general, the experts have two kind of activities. They can evaluate the state of the system to decide if they have any actions to perform or they can carry out the actions they have selected. The former activity does not cause any conflict with other experts, other than competing for computing resources. The latter activity is likely to change the state of the system and affect the actions other experts want to carry out. The scheduler avoids this conflict by assigning priorities to the experts.

To coordinate the activities of the experts the scheduler uses a basic plan which is represented as an augmented transition network. The states in this network are organized hierarchically and represent goals which must be reached to complete the retrieval process. Associated with each state is code that the scheduler uses to monitor the progress of satisfying that goal state. The priorities given to the



The architecture of the expert assistant experts depend on the state in which the system is operating. Once the priorities are established, the scheduler selects which expert will run in the following way. First, all experts that request to evaluate their state are allowed to do so. Those that desire to perform an action signal their intention to do so. Then, the expert with the highest priority is allowed to run.

This algorithm provides more information for the state code than simply running down the priority list evaluating each expert and firing the first one capable of performing an action. With the former algorithm the monitoring code can detect when an expert continually has actions to perform and never gets to do them. This may indicate that the system is not in the proper state and should change.

3.4.2. Anomalous states of knowledge

Belkin, Oddy and Brooks[16] report the results of a design study for an interactive information retrieval system which will determine structural representations of the anomalous states of knowledge (ASKs) underlying information needs, and attempt to rèsolve the anomalies through a variety of retrieval strategies performed on a database of documents represented in compatible structural formats.

The ASK hypothesis is that an information need arises from a recognized anomaly in the user's state of knowledge concerning some topic or situation and that, in general, the user is unable to specify precisely what is needed to resolve that anomaly. Thus for the purposes of information retrieval, it is more suitable to attempt to describe that ASK, than to ask the user to specify his need as a request to the system.

The proposed system would work as follows:

1. The user discusses his information problem in an unstructured statement.
2. The problem statement is converted to a structural representation of the user's anomalous state of knowledge (ASK) by a text analysis program:
   A simple language processor takes the problem statement and identifies the terms and the relationships between them. These relationships may be weak, medium, or strong depending on their proximity in a sentence or paragraph. This information is converted into a graph with the terms as

nodes and the relationships as links. The graph is further reduced by collapsing certain structural features into supernodes. For example, a cluster of terms that are linked by strong links form one kind of supernode and clusters which are linked by medium links form another. The result is a structural representation of a user's information need.

3. According to the type of problem structure, one of several available retrieval mechanisms is chosen to interrogate the database (each member of which is represented by a structural representation of the information associated with the text). In this context, a retrieval mechanism is a strategy for resolving the anomalous aspects of a problem structure. It is accomplished by comparing document and information need representations. It is not necessarily a best match selection in that a document may match part of its structure with the query and be selected for retrieval.
4. The abstract (i.e. the text) is printed out for the user to read. Simultaneously, the user is presented with a brief explanation of why that particular text was chosen (explanation of the retrieval mechanism), indicating aspects of the text structure which the system finds significant in that choice.
5. The system then initiates a structured dialogue with the user, based on the information presented to him, inferring from the response the user's attitude toward: the method of choice; the suitability of the text to the problem; and whether his information need has changed.

6.1. The system changes retrieval mechanism if necessary and/or

6.2. modifies the problem structure if necessary, or

6.3. the system stops if the user is satisfied.

7. The system returns to step 2 or 3.

An interesting result comes from surveying users about how well they thought the graphical representation represented their need and asking authors how well it represented the content of their document abstracts. In general, the people were satisfied with the representations. Some thought a number of terms were too weakly associated; some thought a number were too strongly associated. As these were processed only for a design evaluation, no feedback was taken to refine the representations before using for retrieval. The dissatisfaction with the association strengths, as the authors say, indicates a need for some revision of the association algorithm.

3.4.3. THOMAS

Oddy [17] takes the view expressed by the eminent chemist, Lord Todd: ‘We must surely make the maximum use of computers and associated automation, but if we carry it to the point where the scientist no longer browses in the literature without first formulating questions then I believe we shall do harm to science’.

He introduces a new method of information retrieval by man-machine interaction. The dialogue supported has more symmetry than most interactive computer systems in that the machine forms an image (rather as a man does) of the view of the human enquirer, without requiring him to ask a precise question, and responds with references according to its image.

It is important to try to come to grips with the problem of serving a library user who is not able to formulate a precise query, and yet will recognize what he has been looking for when he sees it. A man, left to his own devices among bookshelves, accomplishes searches of this sort by browsing. During this process, the ‘information need’ tends to be modified, to a greater or lesser extent, by what is found during the search, and the final set of documents, accepted by the searcher as ‘useful’ in relation to his requirements, may be somewhat different in character from the ‘kinds’ of documents he visualized as useful when the search commenced. It is because the information obtained from a document alters the mental state of the reader, that he can conduct this type of browse.

With a good descriptor language, documents which are relevant to a searcher’s problem will have descriptions which he ‘recognizes’ as being promising. The emphasis is on recognition: it is not necessarily true that a query can be formulated in advance by a searcher to match those same descriptions. It seems reasonable to assume that there will be some similarity between the descriptions of documents which are relevant to the same query. But the nature of the similarity may be very subtle and hard to recognize by anybody other than the enquirer. Even so, most approaches to query formulation attempt to predict the description of the required documents. There are occasions when we should not force the searcher to make this prediction, and thus it is that we now recognize that search requirements should be formulated interactively. Oddy states that the enquirer should not be obliged to formulate a question at all.

The program's world model consists solely of knowledge about the organization of literature. Specifically, it is a network of associations between documents, authors, and subject terms: any pair, of like or differing type, may be linked.

To retrieve information, the user introduces a term for a subject related to his need. The first objective of the program, THOMAS, is to find a point in the network knowledge base whose label bears a textual resemblance to the user's term. THOMAS prefers to display a reference immediately, because references are presumed to be what most interest the searcher. If he wishes, a user may mention topics or names without making his decision regarding the document itself.

The program forms an image of the searcher's interest, chooses references for display according to the image and modifies this image continuously in the light of his reactions to the display. A searcher may take as much, or as little, initiative as he wishes while using the program. He may mention new subjects, authors or titles as they come to mind, or he may confine himself to reacting to what he is shown. There are no commands for him to learn to use, such as one usually finds in 'conventional' systems.

However, the program has some disadvantages. After using the program for some time, the user may have a definite idea of what he wants. At this point the model could be used as a formal query, but the program can not take advantage of this; it has no mechanisms for performing any formal search techniques. The only retrieval technique it has is browsing. The network provides enough structure to support the browsing style of retrieval, but it does not capture similarity information which is easily derivable statistically from the documents and terms.

3.4.4. Distributed expert problem treatment

The major function of an information system is the appropriate treatment of the user's problem. By analysing how a general information provision mechanism must operate in order to help the user to treat his problem, a number of discrete functions which interact in complex ways can be identified. This leads Belkin to discuss a particular approach to the modeling and design of problem treatment situations, distributed problem treatment. [18] This approach assumes that problem treatment can be broken down into a number of separate entities, each of which makes hypotheses about its particular area of responsibility, and communicates these results to the other entities of the mechanism.

In project INSTRAT, one of the major problem areas is the architecture of information systems which will not necessarily require a human intermediary. The authors intend to address the issue of systems which help people to solve their problems by providing information or advice relevant to that purpose, and not to any particular type of information system, such as, reference retrieval or fact retrieval.

The minimum functions required of a general information mechanism are:

(i) Problem state: determine the position of the user in the problem treatment process.

(ii) Problem mode: determine the appropriate information provision mechanism; for instance fact retrieval, advice, etc.

(iii) User model: generate a description of the user's type, intentions, beliefs, etc.

(iv) Problem description: generate a description of problem structure, type, topic, context, etc.

(v) Dialogue mode: determine appropriate dialogue mode for immediate situation; for instance use a natural language dialogue or a graphical interaction.

(vi) Relevant world builder: choose and apply appropriate retrieval strategies to the information provision mechanism's world model.

(vii) Response generator: determine the propositional structure of a response to the user which is appropriate to the immediate situation.

(viii) Input analyst: convert input from the user to the mechanism into structures appropriate for the functions.

(ix) Output generator: convert internal response structures into structures appropriate for the user in the immediate situation.

(x) Explanation: describe mechanism operation, restrictions, capabilities, etc. to user at appropriate points in the dialogue.

The basic idea in this approach is that many problems can be usefully decomposed into various elements, each of which can then be treated as a subproblem, each of which is more manageable than the problem as a whole. Distributed problem solving is the general term for this approach, which Belkin modifies to 'distributed problem treatment', to stress that problems are not always solved.

Notice that the components interact with one another in highly complex ways, each depending upon information from all, or almost all the others, at almost all stages of accomplishment of the individual function. This suggests that the traditional, serial structure of systems is inappropriate in the context of the information mechanism.

3.4.5. FASIT, an automatic indexing system

The aim of automatic indexing is to achieve a compact representation of a document suitable for retrieval. FASIT[19] (fully automatic syntactically based indexing of text) identifies content bearing textual units without a full parse, and, without using semantic criteria, groups these units into quasi-synonymous sets.

FASIT is based on the idea that content bearing words or phrases belong to certain syntactic categories or combinations of categories. After assigning the words in the text to categories, it selects concepts based on predefined patterns of categories. It then reduces variations of these concepts to an authoritative form of grouping. The syntactic categories used by FASIT are adapted from an analysis of standard American English. They are based on the traditional eight parts-of-speech in the English language refined by such categories as nominative pronoun, inflected verb, or singular and plural noun. In practice, FASIT indexing consists of two major operations. The first is concept selection and is carried out in three steps:

(i) assignment of words to syntactic categories;

(ii) disambiguation of multiply tagged words; and

(iii) concept selection.

The second is concept grouping and requires two steps:

(i) formation of canonical forms; and

(ii) concept grouping.

Concept selection

1. Assignment of words to syntactic categories

An exception dictionary of words and a suffix dictionary of word endings are used to assign mnemonic tags representing syntactic categories to every word, number, and punctuation character found in the text. Since individual words within the English language may belong to more than one category, more than one tag may be assigned. Any word not tagged by the dictionaries is assigned a default tag of adjective-noun-verb. Examples of tags and their syntactic meanings are:

Syntactic category	Examples
AP (adverb or preposition)	by, around
APP (adverb, preposition, or particle)	in, on, over
GN (general noun)	analysis
JJ (adjective)	administrative
MD (modal auxiliary)	can, may
NN (singular noun)	library
NNS (plural noun)	libraries
PP (preposition)	of, to, from
PPS (singular nominative pronoun)	I, she, he
PQL (prequalifier)	all, half
QL (qualifier)	all, more
SC (subordinating conjunctions)	for, then
VB (uninflected verb)	choose
VBD (past tense verb)	chosen
VBN (past participle verb)	chosen
VBZ (third person singular verb)	chooses

2. Disambiguation of multiply tagged words

Choosing between multiple syntactic categories (disambiguation) is accomplished by examining the tags of words before and after the ambiguous (multitagged) word. For example, the word ‘automated’ may be either a past tense verb or past participle. In the phrase ‘by automated methods’, one rule for disambiguation recognizes that a past tense verb cannot follow ‘by’, a word which functions either as a preposition or adverb, and the past tense tag is removed.

3. Concept selection

The text, represented by tags, is matched against a dictionary of acceptable concepts forms. In the case of ‘by automated methods’, the form identifies ‘automated methods’ as a concept based on the tags ‘past participle’ followed by ‘plural noun’.

These 3 steps are illustrated in the following table where two concept forms are selected as representing the content of the sentence:

<table>
<tr><th colspan="4">A sample of text from a query:</th></tr>
<tr><td colspan="4">I would like all information on library catalogs
produced by automated methods...</td></tr>
<tr><th colspan="4">Tagging and disambiguation (steps 1-2):</th></tr>
<tr><td>Text</td><td>Tag</td><td>Dictionary</td><td>Disambiguated</td></tr>
<tr><td>I</td><td>PPS</td><td>Exception</td><td></td></tr>
<tr><td>would</td><td>MD</td><td>Exception</td><td></td></tr>
<tr><td>like</td><td>VB-SC-JJ</td><td>Exception</td><td>VB</td></tr>
<tr><td>all</td><td>PQL-QL</td><td>Exception</td><td></td></tr>
<tr><td>information</td><td>GN</td><td>Exception</td><td></td></tr>
<tr><td>on</td><td>APP</td><td>Exception</td><td></td></tr>
<tr><td>library</td><td>NN</td><td>Exception</td><td></td></tr>
<tr><td>catalogs</td><td>NNS-VBZ</td><td>Suffix</td><td></td></tr>
<tr><td>produced</td><td>VBD-VBN</td><td>Suffix</td><td></td></tr>
<tr><td>by</td><td>AP</td><td>Exception</td><td></td></tr>
<tr><td>automated</td><td>VBD-VBN</td><td>Suffix</td><td>VBN</td></tr>
<tr><td>methods</td><td>NNS</td><td>Exception</td><td></td></tr>
<tr><th colspan="4">Concept selection (step 3):</th></tr>
<tr><td colspan="2">Concept</td><td colspan="2">Form</td></tr>
<tr><td colspan="2">library catalogs</td><td colspan="2">NN NNS-VBZ</td></tr>
<tr><td colspan="2">automated methods</td><td colspan="2">VBN NNS</td></tr>
</table>

Concept grouping

1. Formation of canonical forms

Each concept is first standardized by purging it of unwanted words, either general nouns, or words such as ‘by’, ‘in’, ‘of’, ‘for’, or ‘to’. For example, ‘of’ is purged from the phrase ‘review of books’. Words for purging are identified by membership in syntactic categories. Categories SC, GN, and PP are examples of categories that are purged. The remaining words of a concept (in stem form) are then sorted. The intent is to merge concepts that differ in minor ways to the same (canonical) form.

2. Concept grouping

Quasi synonymous groups of concepts are formed by treating as equivalent all canonical forms that overlap in at least one stem. In the previous example, the form 'catalog librar' has membership in the group represented by 'catalog' and the group 'librar'.

The following illustrates concept grouping:

Concept	Canonical form	Groups
library catalogs	catalog librar	catalog librar catalog librar
automated methods	autom method	autom method autom method

Tested on a database of 250 documents and 22 queries, FASIT performed better than both thesaurus and stem based indexing systems. Retrievals indicate that the basic idea of FASIT - that significant terms in the text can be identified through syntactic patterns - is valid and that FASIT deserves serious consideration as an advance over stem based systems.

3.4.6. On selection and combining of relevance indicators

In the conventional IRS , each document in the file is characterized by one or more index terms which supposedly describe its content. Those terms are assigned from the natural language or from a pre-prepared list (Thesaurus). Over the years, other means of representing contents were suggested. Also attempts were made to combine several of them assuming independence.

Dealing with the scientific journal literature one can identify several attributes of each item:

(i) journal,

(ii) author(s),

(iii) title and content (which is represented by index terms, manually assigned or automatically derived),

(iv) type (article, review article, conference paper, etc.),

(v) time (when published),

(vi) place (country and language of publication),

(vii) bibliography (items cited in the article),

(viii) citations (items citing the article).

Speaking in general terms, we would like an optimal relevance indicator to be:

(i) exhaustive: all relevant items should be close,

(ii) specific: non-relevant items should not be close,

(iii) objective: the relationship is established by somebody else, other than the author,

(iv) consistent,

(v) time dependent: the indicator should reflect changes in relationships, if and when they occur,

(vi) quality: it is preferable that the attributes selected reflect both closeness and quality,

(vii) available: the citation attribute, for example, is available only quite a long time after the publication date, while the bibliography attribute is available at the time of publication,

(viii) cheap to use: this directly affects efficiency; however, indirectly it might affect effectiveness.

Mansur [20] discusses the attributes of the items in the database and their qualities.

The rationale behind combining relevance indicators is the users' information seeking behavior. Relevance indicators can be combined in several ways. From everyday experience it is known that users are searching by several indicators (authors, citations, index terms, etc.). There is no single indicator which is perfect. Each indicator has some good qualities and some bad ones. In other words, each relevance indicator retrieves several relevant items along with some non-relevant ones (this is called ‘noise’). Those indicators are probably neither totally independent nor highly correlated. They all enhance the relationships among some relevant documents and also enhance, to a certain extent, the relations between relevant and non-relevant ones (noise). If the noise factor is not correlated, i.e. if each indicator links relevant items but different non-relevant ones, then combining indicators allows to enhance relations among the relevant documents while not increasing much of the level of the noise.

3.4.7. The CANSEARCH intermediary system

Pollitt[21] states that there are three particular aspects to expert systems as they relate to information management. Firstly, they provide new mechanisms for capturing information in a very immediate and verifiable form with respect to a collection of knowledge elements. Secondly, the use of an expert system forces a rethink of the methods of organizing and representing information and knowledge in order to make it dynamic and interactive. Finally, the expert system should enable end users to access and question an information collection or knowledge base without requiring them to learn the procedural expressions required of many current systems.

The application of expert systems may be divided into two types. Firstly, as the direct store of knowledge which is required by an end user. Secondly, as the means of access to information or data stored elsewhere, where the capture and update takes place in the ‘remote’ retrieval system.

The CANSEARCH system is a specific example of the second type of application which acts as an intermediary for doctors to aid in their searching for cancer therapy literature.

The knowledge held in this intermediary system can be placed under the following headings:

Subject	Cancer and cancer therapy
Indexing	Medical subject headings Indexing manuals

Retrieval system	MEDLINE/BLAISE
Searching	Recall/Precision/Specificity/Exhaustivity
User	Journal preference Previous searches
Environment	Locally available publications

Knowledge in CANSEARCH is held in a combination of statements and rules.

The statements include an edited subset of the medical subject headings (the controlled vocabulary and thesaurus used by indexers and searchers) which is presented to the searcher for concept and term selection.

Rules are used for the selection and display of frames of concepts and terms, and also for the search statement generation.

3.4.8. A model for an expert system for automated information retrieval

In a searching system for a totally unskilled user it is essential that there is a method of improving the search after an initial response from the database. The method which allows the user to improve the search by adding extra terms chosen from those in the database is one approach, but this is just one enhancement technique which may not be the best tactic. To provide a fully expert system which can assist a user with no information knowledge it is essential that the program can undertake an analysis similar to the intellectual analysis carried out by an intermediary when deciding on refinement strategies.

The searches studied are formulated as a standard Boolean search profile in which a number of synonyms are grouped together in a CONCEPT (Boolean OR statement) and these concepts are combined together to form the full PROFILE using a Boolean AND statement. In the searching system used for the study the Boolean logic and the formulation of the profile in the language of the host computer were carried out automatically by the computer program.

Williams[22] assumes that the user starts with a number of objectives in his search. They are defined as follows:

(i) Pu is the precision required by the user. To simplify the initial analysis this will simply be allowed the values low or high although a more complex system can be defined equally clearly by assuming more levels in the precision requirement.

(ii) Ru is the recall required by the user. Again this is initially allowed only the values high or low.

(iii) Nu is the number of references which the user wants to receive, again with the values high or low.

There are similar variables relating to the output received from the information retrieval system:

(i) Po is the precision of the search output, values high or low.

(ii) Ro is the recall of the search output, values high or low.

(iii) No is the number of references received, values high or low.

Ideally the set (Pu, Ru, Nu) will match the set (Po, Ro, No) or in some cases where the user requirements are exceeded this may be an acceptable result. The set (U) is compared with the set (O) and there are 64 different situations to consider. The cases which occur have been tabulated and the search situation in each case has been studied by experienced searchers to identify the best strategy to improve the search in each of the 64 cases. In the tabulation the high and low values are recorded as + and - respectively. In the comparison table the cases where there is an exact match are recorded as *, those where the search output is higher than the user requirement by + and those where results are lower than the user specification as -. Tables 1 and 2 represent all 64 situations.

Some parameters which define those characteristics of the search profile which are adjusted by the searcher to improve the search. The set of parameters chosen were:

Exhaustivity: The exhaustivity E of a concept in a Boolean search is the number of terms used in any concept.

Table 1. First part of the complete tabulation of possible search situations

	No	Po	Ro	Nu	Pu	Ru	N	P	R
1.	+	+	+	+	+	+	*	*	*
2.	+	+	+	+	+	-	*	*	+
3.	+	+	+	+	-	+	*	+	*
4.	+	+	+	+	-	-	*	+	+
5.	+	+	+	-	+	+	+	*	*
6.	+	+	+	-	+	-	+	*	+
7.	+	+	+	-	-	+	+	+	*
8.	+	+	+	-	-	-	+	+	+
9.	+	+	-	+	+	+	*	*	-
10.	+	+	-	+	+	-	*	*	*
11.	+	+	-	+	-	+	*	+	-
12.	+	+	-	+	-	-	*	+	*
13.	+	+	-	-	+	+	+	*	-
14.	+	+	-	-	+	-	+	*	*
15.	+	+	-	-	-	+	+	+	-
16.	+	+	-	-	-	-	+	+	*
17.	+	-	+	+	+	+	*	-	*
18.	+	-	+	+	+	-	*	-	+
19.	+	-	+	+	-	+	*	*	*
20.	+	-	+	+	-	-	*	*	+
21.	+	-	+	-	+	+	+	-	*
22.	+	-	+	-	+	-	+	-	+
23.	+	-	+	-	-	+	+	*	*
24.	+	-	+	-	-	-	+	*	+
25.	+	-	-	+	+	+	*	-	-
26.	+	-	-	+	+	-	*	-	*
27.	+	-	-	+	-	+	*	*	-
28.	+	-	-	+	-	-	*	*	*
29.	+	-	-	-	+	+	+	-	-
30.	+	-	-	-	+	-	+	-	*
31.	+	-	-	-	-	+	+	*	-
32.	+	-	-	-	-	-	+	*	*

Generality: The generality G of a term describes the extent to which a term embraces a number of aspects or a broad area of knowledge. For example, 'information retrieval' is clearly a very general term whereas 'Boolean searching' is much more specific.

Simplicity: The simplicity S of a search profile refers to the number of concepts linked by the Boolean AND operator.

Ambiguity: The ambiguity A is the degree to which multiple meanings are introduced by using a given term.

Table 2. Last part of the complete tabulation of possible search situations

	No	Po	Ro	Nu	Pu	Ru	N	P	R
33.	-	+	+	+	+	+	-	*	*
34.	-	+	+	+	+	-	-	*	+
35.	-	+	+	+	-	+	-	+	*
36.	-	+	+	+	-	-	-	+	+
37.	-	+	+	-	+	+	*	*	*
38.	-	+	+	-	+	-	*	*	+
39.	-	+	+	-	-	+	*	+	*
40.	-	+	+	-	-	-	*	+	+
41.	-	+	-	+	+	+	-	*	-
42.	-	+	-	+	+	-	-	*	*
43.	-	+	-	+	-	+	-	+	-
44.	-	+	-	+	-	-	-	+	*
45.	-	+	-	-	+	+	*	*	-
46.	-	+	-	-	+	-	*	*	*
47.	-	+	-	-	-	+	*	+	-
48.	-	+	-	-	-	-	*	+	*
49.	-	-	+	+	+	+	-	-	*
50.	-	-	+	+	+	-	-	-	+
51.	-	-	+	+	-	+	-	*	*
52.	-	-	+	+	-	-	-	*	+
53.	-	-	+	-	+	+	*	-	*
54.	-	-	+	-	+	-	*	-	+
55.	-	-	+	-	-	+	*	*	*
56.	-	-	+	-	-	-	*	*	+
57.	-	-	-	+	+	+	-	-	-
58.	-	-	-	+	+	-	-	-	*
59.	-	-	-	+	-	+	-	*	-
60.	-	-	-	+	-	+	-	*	*
61.	-	-	-	-	+	+	*	-	-
62.	-	-	-	-	+	+	*	-	*
63.	-	-	-	-	-	+	*	*	-
64.	-	-	-	-	-	+	*	*	*

No numerical values were ascribed to these parameters. The parameters suggested here are those which are typically controlled by a skilled searcher to improve the performance of the search profile.

By looking at the comparison table of the set (O) and the set (U), the situations where adjustments have to be made can be identified and the appropriate corrections which must be made to the parameters A, E, S, and G can be deduced. These situations have been analysed independently by two search analysts to determine the optimum enhancement strategy in each case. After the model has been refined by discussion with other experts it will provide a mechanism by which a computer driven system can advise the user on the best strategy. It is clear that in some cases the search profile must be modified and in other cases the requirements of the user have to be adjusted to cope with the situation.

The variation of the different parameters can be represented in tables as follows.

The possible actions to improve the search profile are:

Table 3: possible actions to improve the search profile

Increase	No	by increasing	G	S	E
Increase	Ro	by increasing	G	S	E
Increase	Po	by decreasing	A	G	S
Decrease	No	by decreasing	G	S	E
Decrease	Ro	by decreasing	G	S	E
Decrease	Po	by increasing	G	S	

The following table shows how a variation in the control parameters will affect the output of the search.

Table 4: effect of varying A S E and G

Decreasing	A	will increase P	decrease N
Decreasing	S	will increase P	decrease N,R
Decreasing	C	will increase P	decrease N,R
Decreasing	E	will	decrease N,R

By using these tables all of the possible searching situations defined in the model can be studied and appropriate actions to improve a search in each situation can be decided. If the model were perfect and if the parameters could always be correctly calculated then this model would provide a sound foundation for an expert system which would provide adaptive searching for the end-user.

It is assumed that whenever the recall is unsatisfactory the user will be encouraged to increase the exhaustivity E and whenever the precision is poor the user will be pressed to decrease the ambiguity A. This is not mentioned in the following table. There are some situations which arise during searching where the only realistic solution is to consult a skilled intermediary. In order to produce a complete solution from the viewpoint of automated assistance this possibility has been ignored in the table. In the construction of a practical automated search system there will be occasions when the user will be told to give up and consult an expert searcher. This particularly applies when both the precision and the number of items retrieved are poor.

A shorthand representation is used in the table. Where one of the parameters should be reduced to improve the search profile the symbol > is used. Conversely < is used to show that the parameter is to be increased. A ? after a parameter shows that the required corrective action may have an adverse effect on the parameter shown and the overall effect of the modification must be carefully considered.

Table 5: actions to improve the search profile

N	P	R			
+	+	+	Reduce No, limit by user		G>,S>
+	+	*	Reduce No, limit by user	R?	G>,S>
+	+	-	Reduce No, limit by user	R?	G>,S>
+	*	+	Reduce No, limit by user		G>,S>
+	*	*	Reduce No, limit by user	R?	G>,S>
+	*	-	Reduce No, limit by user	R?	G>,S>
+	-	+	Reduce No, increase Po		G>,S>
+	-	*	Reduce No, increase Po	R?	G>,S>
+	-	-	Reduce No, increase Po	R?	G>,S>
*	+	+	Accept		
*	+	*	Accept		
*	+	-	Increase Ro		G<,S<
*	*	+	Accept		
*	*	*	Accept		
*	*	-	Increase Ro	P?	G<,S<
*	-	+	Increase Po		G>,S>
*	-	*	Increase Po	R?	G>,S>
*	-	-	Increase Ro	R?	G>,S>
-	+	+	Accept		
-	+	*	Increase Ro		G<,S<
-	+	-	Increase Ro,No		G<,S<
-	*	+	Accept		
-	*	*	Increase Ro	P?	G<,S<
-	*	-	Increase Ro	P?	G<,S<
-	-	+	Accept		
-	-	*	Accept		
-	-	-	Increase Ro	P?	G<,S<

An example will illustrate the conclusions which can be drawn.

Table 6: example

	No	Po	Ro	Nu	Pu	Ru	N	P	R
9.	+	+	-	+	+	+	*	*	-

In case 9 the user requested a large number of references, a high precision, and a high recall. The results produced the high precision and retrieval numbers but the recall was deficient. The remedy for low recall is to increase the generality G, or increase the simplicity S, or increase the exhaustivity E. The effects of varying these parameters is given in Table 4 and any changes which work against the desired effect must be carefully considered.

3.4.9. Userlink systems

The Userlink system has been built as a production rule system.[15] This means that a processing system called an interpreter has been built which will interact with text files in the form of tables which describe the sequence of actions in the program. Since these are text files they can easily be modified using ordinary text processing techniques. The program steps can therefore be reconfigured simply by editing text files rather than altering a program structure. If new actions are required then new program modules must be created but if it is simply a matter of changing the sequence of actions then only text editing is necessary.

Various types of files are used to drive the interpreter. A rules file is built up which defines the routing of the program from one action to the next. There is a file of configuration details which contain information such as telephone numbers, user authorization codes for the networks, etc. The retrieval language details which provide the words and structure for the search profile are stored. All of the text messages which are displayed to the user are stored in a file so that they can easily be altered as required. There is also a comprehensive file of help messages and tutorial advice to guide the user.

The basic rules file on which the program is based has the following components:

(a) a number which specifies the position reached in the interaction dialogue;

(b) a condition which decides which action will next be carried out;

(c) a piece of text which is displayed to the user;

(d) a list of numbers designating the actions which must be carried out by the program dependent on the condition which has been found to be correct;

(e) a number which defines the next action in the sequence of events following the condition which has just been found.

Thus when the interpreter is at a particular point in the interaction, this is found in the files by searching for the number (a). Those entries which have the correct number in (a) are searched to see whether the condition in (b) is satisfied. There is a unique record with the correct values for (a) and (b). This record will specify which piece of text will be displayed to the user and will indicate a series of actions which the program carries out in (d). Finally the entry in (e) gives the value to be found for the possible situations which can follow at this point. These are looked up in the table to identify the next action in the program.

As well as these rule files there are also files which describe the characteristics of the host computers which are accessed. To change from one host to another it is not normally necessary to do further programming. A generalized description of the search commands on a system has been created and the values for different information retrieval systems are stored. The program has been constructed to incorporate these into a search profile at the time the search is carried out. To access another host, the different expressions of the search commands are inserted in the file and the program then constructs a profile which follows the retrieval language of the new host. The user specific data such as user name and password are also stored in these configuration files.

There is also a set of files which store the variable data which is needed to access the communication system. This will include the telephone number of the network, and the passwords needed by the user.

The result of using these techniques is a system where the adaptation to a new system can often be done by text editing rather than reprogramming and the routing of the program through any new modules which are created is achieved by specifying the conditions in simple tables. The author states that this type of programming is essential in an environment where there is frequent change. It is also a fundamental component of a system which can record the experience of the users and use this knowledge to adapt to the needs of the users.

3.4.10. Idea tactics

Bates[23] performs some more psychological work concerning information retrieval. She describes 17 'idea tactics' which are tactics to help generate new ideas or solutions to problems in information searching. They are intended to help improve specialist's thinking and creative processes in searching. The tactics should be applicable to all kinds of situations: both bibliographic and reference searches, and in both manual and on-line systems.

3.5. Conclusion

At present the end-user, the person with the information need, must often rely on an intermediary to make use of available online systems. Most systems cannot, for example:

(i) respond reasonably to input not conforming to a rigid grammar;

(ii) ask for and understand clarification if the user's input is unclear;

(iii) offer clarification of their own output if the user asks for it.

With the large investment in existing online systems and their successful use by intermediaries, it is difficult to predict how soon a new generation of systems intended for end users will be developed. Fortunately, available technology now allows one to begin to address the need for multiple levels of user interfaces. A remote intelligent terminal at the user site may be programmed to provide a user-oriented interface for a given user or user class. Alternatively, front end processors at the central site or software modules in the main computer system may be used to vary command languages and display formats. Of course, one must use the information and facilities available from the host system which forms a certain limitation.

Researchers in AI are currently exploring systems capable of 'graceful interaction', i.e. dealing appropriately with anything a user happens to say, rather than just those inputs that conform to rigid rules. Without graceful interaction skills, interactive computer systems will continue to appear uncooperative, uncompromising, and altogether obtuse to the non-expert user. Skills required to provide such a capability include:

(i) flexible parsing of elliptical, fragmented and otherwise ungrammatical input;

(ii) an explanation facility;

(iii) focus mechanisms to keep track of what the conversation is about.

Just as the mode of communication is likely to be influenced by AI techniques, the databases to which end users will have access online will also be affected by current work on construction of 'knowledge-based systems'. A new programming methodology has been growing up around the problem of how to transfer human expertise in given domains into machine form, so as to enable computing systems to serve as assistants in the performance of difficult tasks. Steps in the development of a knowledge-based system include:

(i) formulating the application problem;

(ii) designing, constructing and refining a knowledge base of expertise;

(iii) developing schemes of inference, search or problem solving;

(iv) winning the confidence of experts;

(v) evaluating and testing the programs;

(vi) developing production versions of the programs.

4. AN EXPERT SYSTEM FOR PATENT INFORMATION RETRIEVAL

In this section we present a general description of the functionality and architecture of a proposed system that combines state of the art information retrieval and AI technology as well as existing patent information retrieval tools in order to automate the task of a patent information retrieval intermediary.

4.1. Functionality

The emphasis is put on subject searching in patent files, although other types of searches (which seem less complicated) could also be supported.

We believe that subject searching is especially important also if one aims at making patent information directly available for potential users outside the established user community (e.g. small or medium sized companies, scientists, engineers).

At least three methods are available to the user:

IPC hierarchical search

This method takes the user through a series of menus which roughly correspond to the IPC classification scheme. Note that all intricacies and internal rules (e.g. the actual classification codes) of the scheme are transparent to the user. The user can stop the search at any time and switch to another method, the search space of which is confined (in a transparent way) to the current classification group.

Keyword search

Here the user is assisted by the system in the construction of an appropriate search.[2] Obviously, this is an iterative process and the system can employ a number of techniques (e.g. 'zooming') to increase its effectiveness.

It should be noted that this method may also use the IPC classification, although this would not be visible to the user.

Browsing

This important technique[17] is also available: one can browse the search space using a number of criteria to be dynamically selected by the user.

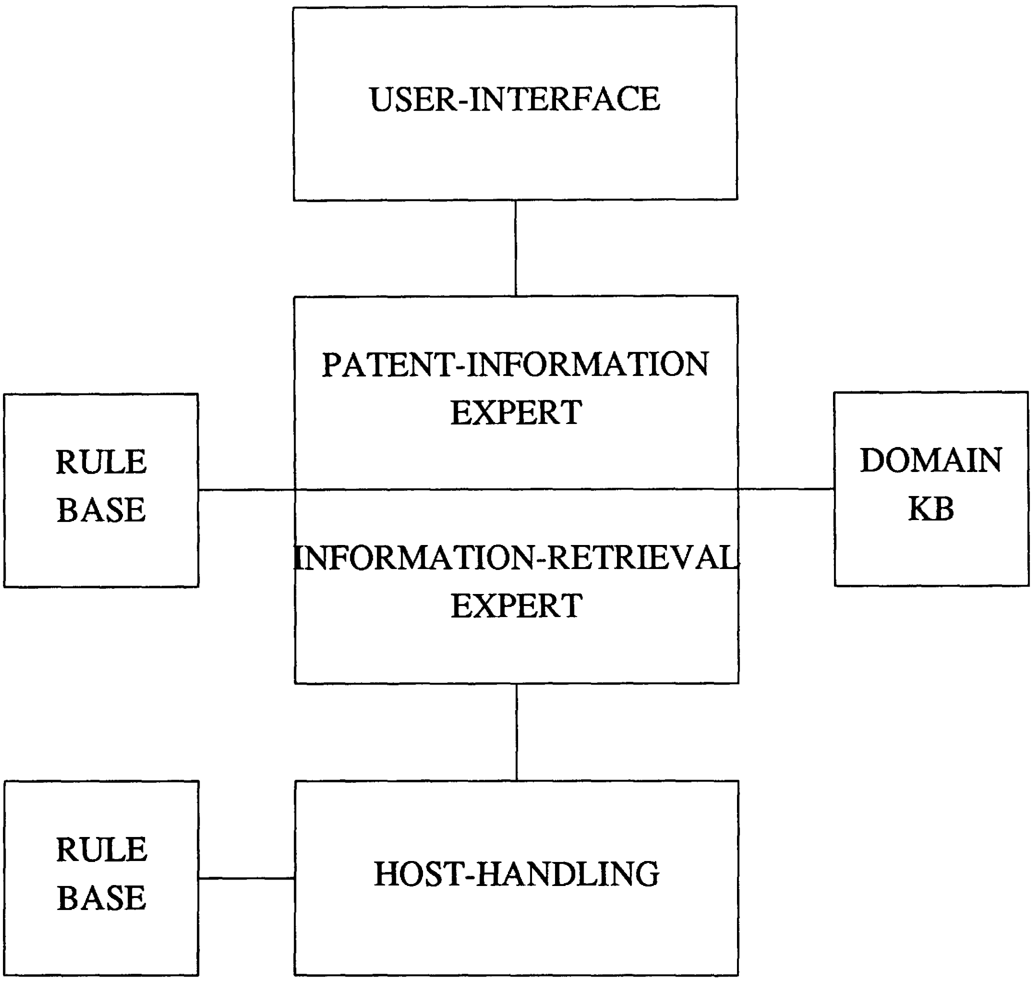
Further functions to be provided by the system include facilities of downloading, reformatting, updating the knowledge base etc.

4.2. User interface

The user interface will be based on graphics, windows and menus, so it will not be necessary for the user to learn a formal command language.

Obviously, comprehensive help facilities should be available throughout the system.

4.3. Architecture



Modules

One can envisage the system as consisting of four layers:

(a) At the lowest level, there is a HOST-HANDLING module that performs relatively low level functions such as:

  (i) managing the connection (dialing, logging on/off, etc.),

  (ii) translating the internal representation of commands to the host's idiom,

  (iii) translating incoming data into the system's internal representation.

(b) Next comes the INFORMATION-RETRIEVAL layer that acts like a general purpose information retrieval expert: it knows about the capabilities of various hosts, how to select a search strategy etc.

(c) Intimately connected to the previous module is the PATENT-INFORMATION specialist layer. It handles problems and strategies that are directly related to the particular structure of patent information. For example, this module can act as a knowledgeable user of the IPC scheme, it knows about patent families etc.

(d) Finally, the top USER-INTERFACE layer is responsible for the presentation of results and the dialogue with the user.

Knowledge bases

Most of the knowledge in the system is not 'hard coded'. Rather it resides in several rule bases (at least one for each module) which can easily be updated.

In addition to these rule bases which can be though of as containing 'procedural' expertise, there are other more fact oriented knowledge bases which take the form of a classical database or a semantic net.

E.g., one such knowledge base represents the IPC hierarchy. Another very important knowledge base is the **domains knowledge base** which consists of a sophisticated thesaurus/semantic network with the following additional characteristics:

(a) It employs a finely grained hierarchy of term relationships, e.g. as described by Larsen.[24]

(b) It is multilingual in order to allow the user to also use his own language in describing concepts.

(c) It contains links to the IPC knowledge base so that the system can make good use of this access scheme, even when browsing or doing a keyword search.

The construction of the initial domain knowledge base could make use of existing tools (e.g. Catchword index). In addition, it could apply automatic indexing techniques (see Section 3) on selected patent information files.

It is envisaged that all knowledge bases (including the rule bases) in the system can be updated by knowledgeable users (e.g. information retrieval experts). The domain knowledge base can be updated by the end-user as well, e.g. to familiarize the system with his own favorite terminology etc.

The update task will be managed by a specialized **knowledge acquisition** module that provides facilities such as consistency checking etc.

Contents of the knowledge base

The list below summarizes some the most important knowledge sources to be represented in the various knowledge bases of the system:

(i) the IPC scheme complemented with extensive additional explanation;

(ii) a keyword list, e.g. the Official Catchword Index, or the Stichwórt Verzeichnis, incorporated into a sophisticated thesaurus (e.g. a semantic network), with pointers to the IPC;

(iii) an inventory of hosts and patent files with their properties;

(iv) the different command languages and retrieval tools of each host;

(v) identification numbers, codes telephone numbers etc.;

(vi) user profiles;

(vii) search profiles;

(viii) a vocabulary of network and host messages with their translation into user friendly statements;

(ix) IPC concordances:

   (a) between the different IPC editions, and

   (b) between the IPC and other subject classifications;

(x) expert knowledge concerning both patent document and online information retrieval;

(xi) the number of patents under each IPC symbol.

4.4. Implementation

Although the system could be implemented as a front-end to the information providing system, we prefer to see it at the user's end in order to obtain more flexibility and information provider independence.

This is especially important in the light of the rapid evolution of mass storage systems which will make it possible in the near future to keep sizable portions of the information bases locally, thus avoiding communication costs.

Hence the system will run on a personal computer or workstation of average capability (e.g. 80286 or 68000 based). As is usual with this kind of programs, there will be a need for a reasonable amount of main memory (say of the order of several megabytes) as well as a large amount of secondary storage to accommodate the initial domain knowledge base (e.g. a CD Rom disk, combined with an updatable hard disk to contain local updates). Even then, it is probably advisable to split the domain knowledge base into separate versions, one for each (sub)discipline.

5. CONCLUSIONS AND RECOMMENDATIONS

In this report we have sketched the importance of patent information bases and the present difficulties which prevent this valuable resource of knowledge to be fully used by all who could benefit from it. Some of the problems with the present situation were uncovered or confirmed in a survey of patent information users, of which we gave a summary of results.

We then presented a state of the art overview of recent developments in software technology which are relevant to the problem of patent information retrieval: namely information retrieval systems, in particular IR expert systems.

Finally, we briefly described a proposal for an expert system that could act as a knowledgeable intermediary between the end user with no information retrieval experience and the various host systems. It appears that the construction of such a system is now feasible, all the more so because it could make use of existing access tools such as the IPC scheme.

During the above mentioned survey, it became clear that there is a great interest in such a system, even among established users of patent information files. However, we believe that the largest benefit to come from it would be the possibility to open up the wealth of technical and scientific information residing in patent information systems to a wider user community who otherwise could not afford the cost of a human intermediary.

We therefore recommend further research into the development of such a system, preferably in cooperation with an academic research institute, a commercial corporation that could later market the resulting software product and an international patent organization (e.g. EPO, WIPO) that could assist in providing part of the expert knowledge to be used by the system.

References

1. D. Sacca, D. Vermeir et al., 'Description of the overall architecture of the KIWI system,' in *Proceedings of the Esprit Technical Week*, 1986. Elsevier Publ. Co.

2. G. Guida and C. Tasso, 'An expert intermediary system for interactive document retrieval ,' *Automatica (GB)*, Vol. 19, No 6, pp. 759-66, Automatica (GB), Nov. 1983.

3. Walker, 'Patents as information - An unused resource,' in *IFLA Journal*, Vol. 10, 1984.

4. Pollick, 'Patents and chemical abstracts service,' in *Science and technology libraries*, Vol. 2, 1981.

5. A. Wittman and R. Schiffels, *Grundslagen der Patentdokumentation*, 1976.

6. Hausser, 'The patent system and the medium-sized industry,' in *World Patent Information*, Vol. 2, 1980.

7. H.L. Shuchman, 'Engineers who patent: data from a recent survey of American bench engineers,' in *World Patent Information (WPI)*, Vol. 5, 1983.

8. *International Patent Classification, Fourth edition 9 volumes, Geneva: WIPO*, 1984.

9. W. Bruce Croft and Roger H. Thompson, 'An expert assistant for document retrieval,' in *COINS Technical Report 85-5* , 1985. Department of Information Science University of Massachusetts Amherst, Massachusetts 01003, USA

10. G. Salton, 'Another look at automatic text-retrieval systems,' in *Communications of the ACM*, Vol. 29, pp. 648-656, 1986.

11. C. Oppenheim, 'The past, present and future of the patents service of Derwent Publications Ltd,' *Science and technology libraries*, Vol. 2, No 2, 1981.

12. K. R. Walton, 'Searchmaster - programmed for the end-user,' in *Online*, pp. 70-79, 1986.

13. Stuart M. Kaback, 'What's in a patent? Information! But can I find it?,' in *Chem. Inf. Comput. Sci.*, Vol. 24, pp. 159-163, 1984.

14. A. Vickery, 'An intelligent interface for online interaction ,' *J. Inf. Sci. Princ. & Pract. (Netherlands)*, Vol. 9, No 1, pp. 7-18, J. Inf. Sci. Princ. &

Pract. (Netherlands), Aug. 1984.

15. P.W. Williams, *The design of an expert system for access to information*, pp. 23-29, 1985.

16. N.J. Belkin, R.N. Oddy, and H.M. Brooks, 'ASK for information retrieval: Part II: Results of a design study,' in *Journal of Documentation* , Vol. 38, pp. 145-164, 1982.

17. R.N. Oddy, 'Information retrieval through man-machine dialogue,' in *Journal of Documentation*, Vol. 33, pp. 1-14, 1977.

18. N. J. Belkin, T. Seeger, and G. Wersig, 'Distributed expert problem treatment as a model for information system analysis and design,' in *Journal of information science*, Vol. 5, pp. 153-167, 1983.

19. M. Dillon and A.S. Gray, 'FASIT: A fully automatic syntactically based indexing system,' in *Journal of the American Society for Information Science*, pp. 99-108, 1983.

20. O. Mansur, 'On selection and combining of relevance indicators,' in *Information Processing and Management*, Vol. 16, pp. 139-153, 1980.

21. A.S. Pollitt, 'A ''front-end'' system: an expert system as an online search intermediary,' in *Aslib Proceedings*, Vol. 36, pp. 229-234, 1984.

22. P.W. Williams, *A model for an expert system for automated information retrieval*, pp. 139-149, 1984.

23. M. J. Bates, 'Idea tactics,' in *Journal of American Society for Information Science*, Vol. 30, pp. 280-289, 1979.

24. H. L. Larsen, *Knowledge representation in IRIS: an information retrieval intermediary system*, 1987. Technical report.

Documents created and delivered electronically: DOCDEL, pp. 124-127.

International Patent Classification, Fourth edition 9 volumes, Geneva: WIPO, 1984.

Patent Information and Documentation Handbook, 5 volumes, (regularly updated), Geneva: WIPO, 1986.

Addis, T.R., 'Expert systems: an evolution in information retrieval,' in *Information technology: Research and development*, pp. 301-324, 1982.

Argentesi, F., L. Constantini, and F. Gardin, 'Design and implementation of a knowledge based system for intelligent information retrieval in a hostile software environment: a software recycling application (NUMSAS/ISIS interface) ,' in *7th Annual Symposium on Safeguards and Nuclear Material Management, Liege, Belgium, 21-23 May 1985*, ed. L. Stanchi, pp. 39-44, Comm. Eur. Commun., Ispra, Italy, 1985.

This paper describes the design and implementation of a knowledge-based system for interfacing a large inconsistent and incomplete database with a statistical package whose input is data retrieved from the data base itself. Due to technical reasons the frame-based shell of the knowledge base system had to be implemented in NATURAL and ADABAS. Solutions adopted for overcoming this apparent limitation are discussed together with a description of tools developed for entering rules and debugging the knowledge base. The particular nature of the knowledge required in the knowledge base and how it was structured represented a key factor for the successful design of the system. This crucial aspect is explained in detail. Performance of such a knowledge-based system is evaluated and an account is given about the time scale of the project. The paradigm of software recycling is introduced, the particular application described in this paper being an example of this paradigm, and the crucial role played by AI in making the recycling process possible is outlined.

Barraclough, E.D., 'On-line searching in information retrieval,' in *Journal of Documentation* , Vol. 33, pp. 220-238, 1977.

The approach taken in this article is to discuss the development of on-line systems, both the computer changes that made it possible and the use made of the changes by information retrieval systems. In this way, by understanding the reasons for the development and for the rejection or acceptance of various techniques, the reader may judge for himself the good and bad points of current systems.

Bates, M. J., 'Idea tactics,' in *Journal of American Society for Information Science*, Vol. 30, pp. 280-289, 1979.

An information search tactic is a move made to further a search. In this article, 17 'idea tactics' are presented: tactics to help generate new ideas or solutions to problems in information searching. The focus of these tactics is psychological; they are intended to help improve the information specialist's thinking and creative processes in searching. The tactics are applicable to all kinds of situations - both bibliographic and reference searches, and in both manual and on-line systems. Research leads for the study of idea tactics are suggested, and experimental design problems associated with the testing of all sorts of search tactics are discussed.

Belkin, N. J., R.N. Oddy, and H.M. Brooks, 'ASK for information retrieval: Part I: Background and theory,' in *Journal of Documentation* , Vol. 38, pp. 61-71, 1982.

We report the results of a design study for an interactive information retrieval system which will determine structural representations of the anomalous states of knowledge (ASKs) underlying information needs, and attempt to resolve the anomalies through a variety of retrieval strategies performed on a database of documents represented in compatible structural formats. Part I discusses the background to the project and the theory underlying it, Part II (next issue) presents our methods, results and conclusions.

Basic premises of the project were: that information needs are not in principle precisely specifiable; that it is possible to elicit problem statements from information system users from which representations of the

ASK underlying the need can be derived; that there are classes of ASKs; and, that all elements of information retrieval systems ought to be based on the user's ASK. We have developed a relatively freeform interview technique for eliciting problem statements, and a statistical word co-occurrence analysis for deriving network representations of the problem statements and abstracts. Structural characteristics of the representations have been used to determine classes of ASKs, and both ASK and information structures have been evaluated by, respectively, users and authors.

Some results are: that interviewing appears to be a satisfactory technique for eliciting problem statements from which ASKs can be determined; that the statistical analysis produces structures which are generally appropriate both for documents and problem statements; that ASKs thus represented can be usefully classified according to their structural characteristics; and, that of 35 subjects, only two had ASKs for which traditional 'best match' retrieval would be intuitively appropriate. The results of the design study indicate that at least some of our premises are reasonable, and that an ASK-based information retrieval system is at least feasible.

Belkin, N.J., R.N. Oddy, and H.M. Brooks, 'ASK for information retrieval: Part II: Results of a design study,' in *Journal of Documentation* , Vol. 38, pp. 145-164, 1982.

We report the results of a design study for an interactive information retrieval system which will determine structural representations of the anomalous states of knowledge (ASKs) underlying information needs, and attempt to resolve the anomalies through a variety of retrieval strategies performed on a database of documents represented in compatible structural formats. Part I (previous issue) discusses the background to the project and the theory underlying it, Part II presents our methods, results and conclusions.

Belkin, N. J., T. Seeger, and G. Wersig, 'Distributed expert problem treatment as a model for information system analysis and design,' in *Journal of information science*, Vol. 5, pp. 153-167, 1983.

By analysing how a general information provision mechanism must operate in order to help the user to treat his problem, we identify a number of discrete functions which interact in complex ways. This leads us to discuss a particular approach to the modeling and design of problem treatment situations, distributed problem treatment. This approach assumes that problem treatment can be broken down into a number of separate entities, each of which makes hypotheses about its particular area of responsibility, and communicates these results to the other entities of the mechanism.

In project INSTRAT, one of the major problem areas is the architecture of information systems which will not necessarily require a human intermediary. The authors intend to address the issue of systems which help people to solve their problems by providing information or advice relevant to that purpose, and not to any particular type of information system, such as, reference retrieval or fact retrieval.

The major function of an information system is the appropriate treatment of the user's problem.

Bose, P.K. and M. Rajinikanth, 'KARMA: knowledge-based assistant to a database system ,' in *Second Conference on Artificial Intelligence Applications: The Engineering of Knowledge-Based Systems (Cat. No 85CH2215-2), Miami Beach, FL, USA, 11-13 Dec. 1985*, pp. 467-72, IEEE Comput. Soc. Press, Washington, DC, USA, 1985.

KARMA is a knowledge based assistant to a relational data base system. The system is designed to aid non-expert users in formulating queries at a conceptual level and guiding them towards the relevant information through a reformulation process. To assist users in retrieving the appropriate data, KARMA maintains a representation of the current partial description of the query.

Chignell, M.H., A. Loewenthal, and P.A. Hancock, 'Intelligent interface design ,' in *IEEE 1985 Proceedings of the International Conference on Cybernetics and Society (Cat. No.85CH2253-3), Tucson, AZ, USA, 12-15 Nov. 1985*, pp. 620-3, IEEE, New York, USA, 1985.

This paper explores the concept of communication between human and non-human intelligent entities. It will focus on the human-machine interface as a translating communication channel. The basic components of an intelligent interface are illustrated using the domain of information retrieval. The intelligent interface method is then applied to the problem of developing a knowledge-based adaptive mechanism where the goal is to control task definition and allocation within a cooperative human-machine system.

Intelligent interfaces are necessary for effective communication between intelligent entities that do not speak the same language. In human-human systems the role of the intelligent interface is carried out by interpreters and intermediaries. As machines become more intelligent entities, steps should be taken to employ intelligent interfaces in human-machine communication. This strategy is preferable to that of forcing either the human or the machine to speak the other language.

Croft, W.B. and L. Lefkowitz, 'Task support in an office system,' in *ACM Transactions on Office Information Systems*, Vol. 2, pp. 197-212, 1984.

A major goal of an office system is to support tasks that are central to office functions. Some office tasks are readily implemented with generic office tools, such as calendars, forms packages and mail. Many tasks, however, involve complex sequences of actions which do not all correspond to tool invocations but, instead, rely on the problem-solving abilities of office workers. In this paper, we describe a system (POISE) that can be used to both automate routine tasks and provide assistance in more complex situations.

Croft, W. Bruce and Roger H. Thompson, 'An expert assistant for document retrieval,' in *COINS Technical Report 85-5* , 1985. Department of Information Science University of Massachusetts Amherst, Massachusetts 01003, USA

The first section discusses the characterization of the problem. The next section reviews recent research directed at increasing the effectiveness and usability of document retrieval systems. Some work in user modeling is examined. The third section summarizes research issues. The last section presents the design of an expert assistant.
The system is an expert assistant that provides information and tools to help a user formulate a query that specifies his information need, and provides a number of search techniques for retrieving documents to meet that need. At the outset of his search, the user may have only a general idea of the information he wants, and may be unsure of how to specify it. The system assists the user by gaining knowledge about him and his need, and uses this knowledge to guide the presentation of information for query refinement. While this interaction occurs, both the user and the system refine and expand what they know. The user gets a better idea of what his query should be; the system gets a better idea of both the user and the need. When information need has been made sufficiently clear, the system will select the most effective search technique to retrieve documents for the user evaluation. The system is more than a conventional retrieval system: it can recognize when the user needs help and offers it, although he is not obliged to accept it. Furthermore, the expert assistant can offer explanations of its actions at a level appropriate to the user.
The system is divided into three major components: the interface manager, the system experts and the knowledge base.

D. Sacca, D. Vermeir et al., 'Description of the overall architecture of the KIWI system,' in *Proceedings of the Esprit Technical Week*, 1986. Elsevier Publ. Co.

Presents an overview of a system to provide user friendly access to heterongeneous databases. The system uses a knowledge representation formalism to represent a unified conceptual view on various information bases. The proposed interface is graphically oriented.

Dalloz, X. and J. Rouget, 'Kunstmatige intelligentie. De ware revolutie?,' in *data DECISIONS*, pp. 15-20, 1986.

Dillon, M. and A.S. Gray, 'FASIT: A fully automatic syntactically based indexing system,' in *Journal of the American Society for Information Science*, pp. 99-108, 1983.

The aim of automatic indexing is to achieve a compact representation of a document suitable for retrieval. FASIT (fully automatic syntactically based indexing of text) identifies content bearing textual units without a full parse, and, without using semantic criteria, groups these units into quasi-synonymous sets.
FASIT is based on the idea that content bearing words or phrases belong to certain syntactic categories or combinations of categories. After assigning the words in the text to categories, it selects concepts based on predefined patterns of categories. It then reduces variations of these concepts to an authoritative form of grouping. The syntactic categories used by FASIT are adapted from an analysis of standard American English. They are based on the traditional eight parts-of-speech in the English language refined by such categories as nominative pronoun, inflected verb, or singular and plural noun. In practice, FASIT indexing consists of two major operations. The first is concept selection and is carried out in three steps:

(i) assignment of words to syntactic categories;
(ii) disambiguation of multiply tagged words; and
(iii) concept selection.
The second is concept grouping and requires two steps:
(i) formation of canonical forms; and
(ii) concept grouping.
Tested on a database of 250 documents and 22 queries, FASIT performed better than both thesaurus and stem based indexing systems. Retrievals indicate that the basic idea of FASIT - that significant terms in the text can be identified through syntactic patterns - is valid and that FASIT deserves serious consideration as an advance over stem based systems.

Dubois, J. E. and Y. Sobel, 'DARC system for documentation and artificial intelligence in chemistry,' *J. Chem. Inf. & Comput. Sci. (USA)*, Vol. 25, No 3, pp. 326-33, J. Chem. Inf. & Comput. Sci. (USA), Aug. 1985.

The DARC system deals with structural information both for documentation and for artificial intelligence endeavors in chemistry. Its topological concepts are briefly reviewed in conjunction with the creative data needs in knowledge information systems (KIPS). Knowledge base, inference engine, and user interface are discussed with reference to the DARC potential in the field of AI and expert systems. AI methodology and its impact on knowledge production are reviewed. New chemical computer-aided design (CAD) tools to develop more creative and innovative research in synthesis planning, structure elucidation, and prediction in drug design are no longer pure prospective challenges.

Guida, G. and C. Tasso, 'An expert intermediary system for interactive document retrieval ,' *Automatica (GB)*, Vol. 19, No 6, pp. 759-66, Automatica (GB), Nov. 1983.

Constructing natural language interfaces to computer systems often requires achievements of advanced reasoning and expert capabilities in addition to basic natural language understanding. In this paper the above issue is tackled in the frame of an actual application concerning the design of a natural language interface for interactive document retrieval.

Harper, D. J. and C. J. Van Rijsbergen, 'An evaluation of feedback in document retrieval using co-occurrence data,' in *Journal of Documentation*, Vol. 34, pp. 189-216, 1978.

This paper reports experiments with a term weighting model incorporating relevance information in which it is assumed that index terms are distributed dependently.

Hausser,, 'The patent system and the medium-sized industry,' in *World Patent Information*, Vol. 2, 1980.

Hawkins, D.T. and L.R. Levy, 'Front end software for online database searching. Part 3: Product selection chart and bibliography,' in *Online*, pp. 49-58, May 1986.

This article presents a selection chart containing data on front ends and gateways known to us as of late 1985. It also includes a bibliography of published articles on these products.

Hjerppe, R., *What artificial intelligence can, could, and can't do for libraries and information services*, pp. 7-25, 1983.

Husby, O., A. Midtun, I. Solvberg, and A. Aamodt, *Expert systems in libraries. Toys or tools? Threat or help?*, University of Trondheim - Norway, 1985.

Ingwersen, P., 'Information technology-which applications? ,' *Soc. Sci. Inf. Stud. (GB)*, Vol. 4, No 2-3, pp. 185-96, Soc. Sci. Inf. Stud. (GB), April-July 1984.

After a short discussion of selected characteristics of the information technology (IT) in operation at present, the paper emphasizes and discusses four different views on IT: a denial, a goal, a utility and a craft approach. This is followed by a critical examination of three selected areas to which IT is applied. Electronic data generation and dissemination, or information handling, is exemplified by means of office automation and electronic publishing. Advantages, problems and disadvantages connected to the selected areas are outlined.
New concepts, such as 'downloading' from mainframes to micros and 'intelligent knowledge-based sys-

tems', are considered in relation to data dissemination and information retrieval. Today's IT is found to be a valuable support to the experienced customer but less value to, for example, casual searchers and other less experienced user groups. The paper points to three different approaches applied by the information industry to solve some of the problem areas raised - in particular related to man-system interface: the traditional, the instructive and the automated approach.

Jones, K.P., *Searching techniques: time to replace intelligent operators by intelligent systems*, pp. 49-59, 1982.

Jones, K. Sparck, 'Intelligent retrieval ,' in *Proceedings of Informatics 7*, pp. 136-142, London, 1983.

It is well worth doing research on how to build expert systems. This would not represent a retreat from the glorious days of fully-automatic statistically-based searching. It would recognize that the user has something to contribute, and that cooperation between system and user can be useful. It would not necessarily preclude the use of statistical techniques: these could supply some document-based information to enhance the subject-oriented knowledge of the front end.

Kaback, Stuart M., 'What's in a patent? Information! But can I find it?,' in *Chem. Inf. Comput. Sci.*, Vol. 24, pp. 159-163, 1984.

The information in chemical patents is used for many purposes, including but not limited to patentability, validity, and infringement studies, state of the art reviews, and the monitoring of competitive technology. For some of these purposes, only the information in patent claims is germane; for others, the examples are most important. Information is sometimes presented in a highly specific fashion, while at other times it is quite generic. Searches similarly are sometimes aimed at specifics and other times at generics - and often a specific search must contend with generic patents and vice versa. Overlying these problems are the difficulties of dealing with pictorial and numerical information and with the context in which information is presented. This paper examines these problems, taking a close look at the kinds of information needed by the users of patent information.

Kehoe, C.A., 'Interfaces and expert systems for online retrieval,' in *Online review*, Vol. 9, pp. 489-505, 1985.

This paper reviews the history of separate online system interfaces, leading to efforts to develop expert systems for searching databases, particularly for end users, and introduces the research in such expert systems. Appended is a bibliography of sources on interfaces and expert systems for online retrieval.

Krawczak, D.A., P. J. Smith, S. J. Shute, and M. Chignell, 'EP-X: a knowledge-based system to aid in searches of the environmental pollution literature ,' in *Second Conference on Artificial Intelligence Applications: the engineering of knowledge-based systems (Cat. No.85CH2215-2), Miami Beach, FL, USA, 11-13 Dec. 1985*, pp. 552-7, IEEE Comput. Soc. Press, Washington, DC, USA, 1985.

EP-X (environmental pollution expert) is a prototype expert system that acts as an expert search intermediary for a bibliographic information retrieval system. It searches for documents in the chemical abstracts database in the domain of environmental pollution, serving as an intelligent user interface that accommodates a wide variety of user backgrounds and interests. This paper discusses the general goals of document retrieval for different users, the types of knowledge necessary to accomplish these goals, and some specific examples of how EP-X uses its knowledge.

Larsen, H. L., *Knowledge representation in IRIS: an information retrieval intermediary system*, 1987. Technical report.

Describes the representation and utilization of the domain knowledge needed by IRIS. The main features are the representation of knowledge in the OOPS language, and the use of term relationships. IRIS is a part of the KIWI project.

Lebowitz, M., 'Intelligent information systems ,' in *Proceedings of the Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Bethesda, MD, USA, 6-8 June 1983*, ed. J.J. Kuehn, pp. 25-30, ACM, Baltimore, MD, USA, 1983.

Natural language processing techniques developed for artificial intelligence programs can aid in constructing powerful information retrieval systems in at least two areas. Automatic construction of new concepts

allows a large body of information to be organized compactly and in a manner that allows a wide range of queries to be answered. Also, using natural language processing techniques to conceptually analyze the documents being stored in a system greatly expands the effectiveness of queries about given pieces of text. However, only robust conceptual analysis methods are adequate for such systems. This paper discusses approaches to both concept learning, in the form of generalization-based memory and, powerful, robust text processing achieved by memory-based understanding. These techniques have been implemented in the computer IPP system, a program that reads, remembers and generalizes from news stories about terrorism, and RESEARCHER, currently in the prototype stage, that operates in a very different domain (technical texts, patent abstracts in particular).

Levine, E.H. and J.K. Burnam, *Low cost enhancement of online searching and information center support with an intelligent floppy disk*, pp. 359-363, 1982.

Levy, L.R. and D.T. Hawkins, 'Front end software for online database searching. Part 2: The marketplace.,' in *Online*, pp. 33-40, Jan. 1986.

Online database producers and retrieval services (databanks) are currently in a period of intense change because their traditional market, professional search intermediaries, is close to saturation. The present online information industry growth rate - as much as 30% per year - must be sustained for the participants to remain profitable; the most obvious way is to reach out into virgin territory and expand into new markets. Current efforts to reach end-users of information are a clear indication of this trend; front end software is a prime example of a product for the end-user. This article analyzes the front end software marketplace and discusses some of the complex forces influencing it.

Lith, P. van, 'Ze zijn niet intelligent, ze redeneren alleen,' in *AG- report*, pp. 32-35, 1985.

Mansur, O., 'On selection and combining of relevance indicators,' in *Information Processing and Management*, Vol. 16, pp. 139-153, 1980.

In the conventional IRS , each document in the file is characterized by one or more index terms which supposedly describe its content. Those terms are assigned from the natural language or from a pre-prepared list (Thesaurus). Over the years, other means of representing contents were suggested. Also attempts were made to combine several of them assuming independence.
This paper discusses the attributes of the items in the database and their qualities. It seems that there is no single one which has all the desired qualities. If the attributes are not totally independent neither highly correlated then combining them in a certain way may increase effectiveness. The justification for this comes from the users' information seeking behavior: users are using index terms, author's names, citations, and other attributes in their searches.
A model to accommodate the above hypothesis is formulated and the small experiment performed indicates that the hypothesis may be true, and this way of combining might improve effectiveness.

Martin, W.A., *Helping the less experienced user*, pp. 67-76, 1982.

Neches, R., B. Balzer, N. Goldman, and D. Wile, 'Transferring users' responsibilities to a system: the information management computing environment ,' in *INTERACT '84. First IFIP Conference on 'Human-Computer Interaction', London, England, 4-7 Sept. 1984*, Vol. 1, pp. 195-200, Elsevier, Amsterdam, Netherlands, 1984.

User difficulties in developing and using a useful system model appear to depend both on the number and complexity of rules they must learn, and the memory load of building and applying the required set of rules. The information management computing environment, by presenting several non-traditional capabilities based on a fusion of AI and database technology, may help ameliorate some of these difficulties. The key concepts of the system consist of a uniform underlying database representing all information in the environment, support for retrieval of information objects by description and the explicit specification of rules telling the system how to take over responsibility for consistency maintenance and routine user actions.

Noerr, P.L. and K.T. Bivins Noerr, *The display device: a user-oriented intelligent terminal*, pp. 373-378, 1982.

Oddy, R.N., 'Information retrieval through man-machine dialogue,' in *Journal of Documentation*, Vol. 33, pp. 1-14, 1977.

Oddy introduces a new method of information retrieval by man-machine interaction. The dialogue supported has more symmetry than most interactive computer systems in that the machine forms an image (rather as a man does) of the view of the human enquirer, without requiring him to ask a precise question, and responds with references according to its image.
It is important to try to come to grips with the problem of serving a library user who is not able to formulate a precise query, and yet will recognize what he has been looking for when he sees it. A man, left to his own devices among bookshelves, accomplishes searches of this sort by browsing. During this process, the 'information need' tends to be modified, to a greater or lesser extent, by what is found during the search, and the final set of documents, accepted by the searcher as 'useful' in relation to his requirements, may be somewhat different in character from the 'kinds' of documents he visualized as useful when the search commenced.
It is because the information obtained from a document alters the mental state of the reader, that he can conduct the type of browse described above.

Oppenheim, C., *Patent information online: a review*, p. 91,100.

Patents are an important and under-rated source of information. Some of the features which make patent databases unique are described. Some criteria for the evaluation of online patents databases are then discussed and the major databases are evaluated on the basis of these criteria. The likely impact of optical discs on patents information retrieval is discussed, and this is followed by some predictions about the future of patents information online.

Oppenheim, C., 'The past, present and future of the patents service of Derwent Publications Ltd,' *Science and technology libraries*, Vol. 2, No 2, 1981.

Ornager, S. and M. Johne, *Changes in thesaurus construction caused by the use of boolean searching*, pp. 167-173, 1983.

Pollick,, 'Patents and chemical abstracts service,' in *Science and technology libraries*, Vol. 2, 1981.

Pollitt, A.S., *An expert system as an online search intermediary*, pp. 25-32.

Improvements can be made in the use of existing online search systems by using techniques developed in work on AI. One particular area receiving attention is the expert system and the use of such a system interposed between a user and the search system in place of the human intermediary. This system could use knowledge of the search system, the subject and search strategies to act in a facilitating, advisory and explanatory role. An ambitious system may even involve some conversion from natural language to the controlled syntax of a command language. The construction of an expert system to act as a local intermediary for searches concerned with cancer therapy is being considered, some of the features to be provided are identified and related to work on expert systems with a view to directing the design and implementation of an improved end user interface to relevant databases.
The knowledge for the proposed expert system may belong to four categories:
1. System knowledge: The command language and facilities available in the search system(s) from logging-on and the submission of search statements to the printing of references or abstracts.
2. Searching knowledge: Relating to the strategy and tactics to be employed in searching.
3. Subject Knowledge: Particular knowledge which can express typical searches for Cancer Therapy, given a thesaurus of terms as a linking framework between the user and the information being sought. The applicability and appropriateness of index terms taken from the guides for indexers is knowledge which refines the vehicle of the thesaurus being knowledge as to how and why these terms are applied.
4. User knowledge: Knowledge about each individual user including previous searches, preferred journals and personal reference collection.

Pollitt, A.S., 'A "front-end" system: an expert system as an online search intermediary,' in *Aslib Proceedings*, Vol. 36, pp. 229-234, 1984.

There are three particular aspects to expert systems as they relate to information management. Firstly, they provide new mechanisms for capturing information in a very immediate and verifiable form with respect to a collection of knowledge elements. Secondly, the use of an expert system forces a rethink of the methods of organizing and representing information and knowledge in order to make it dynamic and interactive. Finally, the expert system should enable end users to access and question an information collection or knowledge base without requiring them to learn the procedural expressions required of many current systems.

The application of expert systems may be divided into two types. Firstly, as the direct store of knowledge which is required by an end user. Secondly, as the means of access to information or data stored elsewhere, where the capture and update takes place in the 'remote' retrieval system.

The CANSEARCH system is a specific example of the second type of application which acts as an intermediary for doctors to aid in their searching for cancer therapy literature.

Rich, E., 'Artificial intelligence and the humanities,' in *Computers and the humanities*, Vol. 19, pp. 117-122, 1985.

Rowe, N.C., 'Modeling degrees of item interest for a general database query system,' *Int. J. Man-Mach. Stud. (GB)*, Vol. 20, No 5, pp. 421-43, Int. J. Man-Mach. Stud. (GB), May 1984.

Many databases support decision-making. Often this means choosing between alternatives according to partly subjective or conflicting criteria. Database query languages are generally designed for precise, logical specification of the data of interest, and tend to be awkward in these circumstances. Information retrieval research suggests several solutions, but there are obstacles to generalizing these ideas to most databases.

To address this problem we propose a methodology for automatically deriving and monitoring 'degrees of interest' among alternatives for a user of a database system. This includes (a) a decision theory model of the value of information to the user, and (b) inference mechanisms, based in part on ideas from AI, that can tune the model to observed user behavior. This theory has important applications to improving efficiency and cooperativeness of the interface between decision-maker and a database system. We have performed some preliminary experiments with it.

Runit,, 'Expert systems - A short introduction,' in *NIF-Seminar: Applications for expert systems in offshore related industry*, 1985.

Salton, G., 'Another look at automatic text-retrieval systems,' in *Communications of the ACM*, Vol. 29, pp. 648-656, 1986.

The effectiveness of a retrieval system is usually evaluated in terms of a pair of measures, known as recall and precision. Recall is the proportion of relevant material actually retrieved from a file, while precision is the proportion of the retrieved material that is found to be relevant to the user's needs. In principle, a search should achieve high recall by retrieving almost everything that is relevant, while at the same time maintaining high precision by rejecting a large proportion of extraneous items. In practice, it is known that recall and precision tend to vary inversely, and that it is difficult to retrieve everything that is wanted while also rejecting everything that is unwanted. A very specific query formulation produces high precision and hence low recall performance. As the query formulation is broadened, more relevant items are retrieved, thus improving the recall, but also more nonrelevant ones, thereby depressing the precision.

Shenton, K., *Graphic retrieval of patent information*, pp. 43-59.

Since 1963, Derwent has been involved in the retrieval of information in chemical patents. Initially, a punch card coding system, based on chemical fragments, was used. This system was modified first for use on mainframe computers, and was later adapted for retrieval online. However, fragmentation coding systems do have the inherent disadvantage of retrieving false drops due to the inability to represent the topology of the indexed structures. Derwent has therefore embarked on an ambitious research program into the graphics indexing of Markush (generic) structures.

This paper summarizes the present position in two areas:

1. Graphic input of Markush structures, including development of software of input and search, and indexing of Markush structures;
2. Translation of graphic input into the Derwent fragmentation code.

Shoval, P., 'Expert/consultation system for retrieval data-base with semantic network of concepts ,' in *SIGIR Forum (USA), Proceedings of the Fourth International Conference on Information Storage and Retrieval, Oakland, CA, USA, 31 May-2 June 1981*, Vol. 16, pp. 145-9, Summer 1981.

This paper describes a development and implementation of an expert-consultation system for a retrieval data-base, that interfaces between the user and a retrieval system. The system's objective is to perform the information consultant's job in assisting a user to select the right vocabulary terms for his query. It is particularly useful for a novice user of a controlled-vocabulary, index-based retrieval system, who is not familiar with the vocabulary and the system Thesaurus. The user will enter his terms/keywords, that represent his information need, and the system will apply search procedures on its knowledge base, and will find relevant concepts to be used as query terms. The system is interactive; it can explain to the user why and how a concept was discovered or suggested, and it can backtrack and try to find alternatives in case the user rejects a suggested concept. Two versions of the system were developed, utilizing two search and interaction strategies.
The expert system
A computerized expert/consultation system ought to perform the job of a human expert. A human expert has the knowledge in the subject area (whether he remembers or has access to knowledge) and working methods/procedures that he applies on the knowledge in order to find the solution for a given problem.
The knowledge is represented as a semantic network; where nodes are terms in the subject area, and links are the various types of relationships between them. The second component of the expert system is the procedures/search algorithms.

Shuchman, H.L., 'Engineers who patent: data from a recent survey of American bench engineers,' in *World Patent Information (WPI)*, Vol. 5, 1983.

Smith, L.C., 'Implications of artificial intelligence for end user use of online systems ,' *Online Rev. (GB)*, Vol. 4, No 4, pp. 383-91, Online Rev. (GB), Dec. 1980.

This paper reviews a number of studies which demonstrate how AI techniques can be applied in the design of end user-oriented interfaces to existing online systems as well as in the development of future generations of online systems intended for the end user.
At present the end-user, the person with the information need, must often rely on an intermediary to make use of available online systems. Most systems cannot, for example:
(i) respond reasonably to input not conforming to a rigid grammar;
(ii) ask for and understand clarification if the user's input is unclear;
(iii) offer clarification of their own output if the user asks for it.
Some criteria for user-oriented online information retrieval systems are:
(i) the language of a system should be easy to understand;
(ii) transactions with a system should be courteous;
(iii) a system should be quick to react;
(iv) a system should relieve the users of unnecessary chores;
(v) be capable of being queried in English language form without the need to use either controlled terms or formal Boolean search logic;
(vi) require minimum of keybording and compensate for common types of error;
(vii) be capable of weighting search terms automatically and ranking documents by degree of match with the search statement;
(viii) provide assistance to the user by guiding him along paths likely to lead to relevant documents;
(ix) be capable of being searched in ways that are unconventional but convenient for the user, e.g. once the user has located some relevant items he should be able to instruct the system to find others 'like them'.
Interim solutions
With the large investment in existing online systems and their successful use by intermediaries, it is difficult to predict how soon a new generation of systems intended for end users will be developed. Fortunately, available technology now allows one to begin to address the need for multiple levels of user inter-

faces. A remote intelligent terminal at the user site may be programmed to provide a user-oriented interface for a given user or user class. Alternatively, front end processors at the central site or software modules in the main computer system may be used to vary command languages and display formats. Of course, one must use the information and facilities available from the host system which forms a certain limitation.

The incorporation of relevance feedback acknowledges the fact that document retrieval is a trial-end-error process.

New approaches to man-machine communication

Researchers in AI are currently exploring systems capable of 'graceful interaction', i.e. dealing appropriately with anything a user happens to say, rather than just those inputs that conform to rigid rules. Without graceful interaction skills, interactive computer systems will continue to appear uncooperative, uncompromising, and altogether obtuse to the non-expert user. Skills required to provide such a capability include:

(i) flexible parsing of elliptical, fragmented and otherwise ungrammatical input;
(ii) an explanation facility;
(iii) focus mechanisms to keep track of what the conversation is about.

Knowledge-based systems

Just as the mode of communication is likely to be influenced by AI techniques, the databases to which end users will have access online will also be affected by current work on construction of 'knowledge-based systems'. A new programming methodology has been growing up around the problem of how to transfer human expertise in given domains into machine form, so as to enable computing systems to serve as assistants in the performance of difficult tasks. Steps in the development of a knowledge-based system include:

(i) formulating the application problem;
(ii) designing, constructing and refining a knowledge base of expertise;
(iii) developing schemes of inference, search or problem solving;
(iv) winning the confidence of experts;
(v) evaluating and testing the programs;
(vi) developing production versions of the programs.

What is particular interesting about such systems is that there are three different modes of end user use in contrast to the single mode for conventional information retrieval systems:

(i) user as a client: get answers to problems;
(ii) user as tutor: improve the systems knowledge;
(iii) user as pupil: harvest the knowledge base.

Smith, Linda C., 'Artificial intelligence in information retrieval systems,' in *Information processing and management*, Vol. 12, pp. 189-222, 1976.

Smith, Linda C., 'Artificial intelligence applications in information systems,' in *Annual review of information science and technology*, Vol. 15, pp. 67-105, 1980.

Practical reasons for looking at AI in the context of computer-based information systems include the possibility of making systems accessible to a wider range of people, delegating certain tasks to the system while helping the user with more complex tasks.

Four AI concepts have particular significance for information systems: pattern recognition, problem solving, representation and learning.

Pattern recognition is the identification of an object with a particular set of features as the member of some class. The problem of reference retrieval is similar to that of pattern recognition. In reference retrieval the development of document surrogates and query formulations may be viewed as a classification problem. The term 'features' is a useful generic term in the context of reference retrieval as well as AI, for it permits one to think of index terms, authors, citations, etc. all as possible features. Feature selection can occur in retrieval systems at two points: when document surrogates are prepared and when queries are formulated for comparison with document surrogates. Items in the file are classified in response ro each query, when the portion to be retrieved is separated from that which is not. Although sorting as an approach to classification requires that the user query to find all items 'like' one already known to the user is a problem of prototype matching. The system assesses the probable relevance of a document to a query by calculating a measure of similarity between a document surrogate and the query formulation. An item is retrieved if the similarity measure is above some threshold.

The representation is a formalism for the knowledge possessed by a system. It may be thought of as 'a set of conventions about how to describe things'. Just as representation in AI is a formalism for knowledge possessed by a system, a document representation is "a formalized statement of the nature of a document". A query formulation may be viewed similarly. When one thinks of computer-based systems rather than manual, one must ask how to take the available information and represent it in a way that the computer can store and manipulate. This includes not only representations for documents and queries, but also relations between documents (such as citation relations) and between terms (such as those shown by a thesaurus). Online systems must be designed to encompass not only internal representations, i.e. representations of information within the computer subsystem, but also external representations - the displays of information at the user-computer interface.

Problem solving is the art of using knowledge effectively to attain desired goals. Problem solving can be approached using either algorithms or heuristics. In reference or data retrieval, the problem confronting the system is to identify, in response to each query, the portion of the contents of the file which should be retrieved. In the case problem solving includes development of a search strategy and the use of some inference mechanism. Application of heuristics could be the use of techniques which allow one to quickly select the subset of the file satisfying the query. Online systems must be designed to include consideration of how best to build the user-computer interface so that poorly constructed queries can be converted to well-structured forms that the computer subsystem can handle.

System elements such as feature selections routines, representations, and heuristics are of course all initially selected and programmed by a human designer when an AI system is developed for some application. Learning mechanisms by which a system can improve its performance over time are therefore necessary so that the initial design does not circumscribe system capabilities. The availability of online computer systems makes it reasonable to speak about dynamic systems which change and improve performance over time. Learning in retrieval systems can have either short term or long term effects. Short term learning is the modification of system response during the processing of a particular query in order to better meet the needs of the user. In reference retrieval systems, for example, this can be done through feedback in query processing, taking account of the relevance status of a sample of retrieved documents as fudged by the user. Long term learning could involve modifying and/or extending the representation to improve system response over time. Modification can include changes in file organization and in item representation, e.g., updating the database to reflect new terminology. Extension can include techniques for storing previous search strategies in a form suitable for subsequent use by other system users.

Sowizral, H.A., 'Expert systems,' in *Annual Review of Information Science and Technology*, Vol. 20, 1985.

Thompson, R.H., 'An expert system for document retrieval ,' in *Expert Systems in Government Symposium (Cat. No.85CH2225-1), McLean, VA, USA, 24-25 Oct. 1985*, ed. Karna, K.N. Croft, pp. 448-56, IEEE Comput Soc. Press, Washington, DC, USA, 1985.

Current experimental information retrieval systems use statistical methods for indexing and retrieval. While these methods have advantages in both their efficiency and effectiveness, they only use superficial knowledge of the users and their information need. Furthermore, they are generally limited to a single retrieval technique. In this paper we describe the design of an expert assistant for document retrieval. The main components of the system are a collection of function specific experts, a knowledge base, and an interface manager. The experts construct detailed models of the user and the information need via interaction with the user and the knowledge base. Based on these models, the system's activity is coordinated by a scheduler using plans, which are derived from analysis of end-user / search intermediary interaction.

Tittlbach, G., *Online patent information with text and graphics via STN international*, pp. 95-104.

This paper discusses current and short-dated planned online patent information services offered via the STN Node Karlsruhe. Special attention is addressed to ongoing developments in extending the access to the German Patent Database PATDPA containing both text and graphical representations. The novel feature is the joint storage of text and drawings in one database, the conversion of digitized graphical data into vectorgraphics output format and the combined transmission of text and graphics via telecommunication networks to various types of terminals.

Vermeir, D., 'OOPS: a knowledge representation language.,' in *Proceedings of the 1986 IEEE International Conference on Computer Languages*, 1986.

Describes OOPS, a knowledge representation language that was developed within the framework of the KIWI project. OOPS is used to represent domain and other knowledge needed to provide an intelligent user friendly interface to information bases.

Vickery, A., 'An intelligent interface for online interaction ,' *J. Inf. Sci. Princ. & Pract. (Netherlands)*, Vol. 9, No 1, pp. 7-18, J. Inf. Sci. Princ. & Pract. (Netherlands), Aug. 1984.

In this paper, the author discusses the ways of improving the performance of online retrieval systems by introducing an automated interface between the enquirer and the system. In the first part of the paper, the main features of such human/machine interaction and the characteristics that the user would like to see incorporated in an interface, are particularly relevant to the problems of implementing an intelligent interface, are discussed. The author concludes with a summary of automated mechanisms that will be needed to improve the quality of interaction between the user and the search system.

Walker,, 'Patents as information - An unused resource,' in *IFLA Journal*, Vol. 10, 1984.

Walton, K. R., 'Searchmaster - programmed for the end-user,' in *Online*, pp. 70-79, 1986.

The searchmaster programs have proven to be both popular and useful, permitting end-users with no previous experience to perform routine searches online with virtually no assistance. At the same time, it has relieved the burden of routine searching for the Information Center's searching staff allowing them to focus on more sophisticated information problems. According to Walton, the searchmaster programs does not represent the ultimate in online search packages for end-users at Exxon. Instead, he simply states that the searchmaster programs perform a valuable service today and offer a foundation for designing and evaluating improved systems tomorrow.

Wilensky, R., Y. Arens, and D. Chin, 'Talking to UNIX in English: an overview of UC,' in *Communications of the ACM*, Vol. 27, 1984.

UC is a natural language help facility which advises users in using the UNIX operating system. Users can query UC about how to do things, command names and formats, online definitions of UNIX or general operating systems terminology, and debugging problems in using commands.

Williams, P.W., *A model for an expert system for automated information retrieval*, pp. 139-149, 1984.

In the course of this work an analysis was undertaken of the actions which a skilled intermediary needs to take to improve a search strategy in the light of feedback from the database records. By formalizing the factors which influence search refinement, a model has been constructed which completely specifies all possible search situations. The possible responses in each of these situations were studied by two skilled intermediaries and the best strategies were defined. This provides the knowledge base on which a complete expert system can be built.

Williams, P.W., *The design of an expert system for access to information*, pp. 23-29, 1985.

This paper is concerned with the construction of programs to provide information access for end users. The programs must provide expert guidance in extracting the desired information and must be designed to adapt to the changing of information systems and communications networks.
An expert system has the following components:
1. A knowledge base which contains important information in the subject area and the connections between the different pieces of information.
2. An inference mechanism which is able to use the connections between the information to make conclusions or formulate advice to present to the user.
3. An explanation system which tells the user why certain actions were taken.
4. A system for adapting the inference mechanism to the experience gained from recording the user activity.
The expert system for information retrieval is required to mimic the capabilities of a skilled information professional. This involves many different facets.

1. Choosing a database which contains the required information.
2. Choosing retrieval terms which contains the required information.
3. Knowledge of the retrieval language commands.
4. Knowledge of the communications systems and protocols.
5. Ability to interact with the host computer dialogue.
6. Ability to react to error messages.
7. Ability to modify the search in the light of the results obtained from the information retrieval system.

Wittman, A. and R. Schiffels, *Grundslagen der Patentdokumentation*, 1976.

Wolpert, S.A., *User friendly systems*, pp. 147-155, 1983.

European Communities — Commission

EUR 11326 — The application of recent software technology to the access to patent information systems

D. Vermeir, E. Laenens, J. Dierick

Luxembourg: Office for Official Publications of the European Communities

1988 — IV, 265 pp., 10 tab., 2 fig. — 21.0 × 29.7 cm

Information management series

EN

ISBN 92-825-7773-2

Catalogue number: CD-NA-11326-EN-C

Price (excluding VAT) in Luxembourg
ECU 7 BFR 300 IRL 5.50 UKL 5 USD 8.50

This report presents an overview of recent developments in software technology, especially information retrieval and expert systems. Particular consideration is given on the possible applications in the area of user-friendly access to patent information systems.

Venta y suscripciones · Salg og abonnement · Verkauf und Abonnement · Πωλήσεις και συνδρομές
Sales and subscriptions · Vente et abonnements · Vendita e abbonamenti
Verkoop en abonnementen · Venda e assinaturas

BELGIQUE / BELGIË

Moniteur belge / Belgisch Staatsblad
Rue de Louvain 40-42 / Leuvensestraat 40-42
1000 Bruxelles / 1000 Brussel
Tél. 512 00 26
CCP / Postrekening 000-2005502-27

Sous-dépôts / Agentschappen:

Librairie européenne /
Europese Boekhandel
Rue de la Loi 244 / Wetstraat 244
1040 Bruxelles / 1040 Brussel

CREDOC
Rue de la Montagne 34 / Bergstraat 34
Bte 11 / Bus 11
1000 Bruxelles / 1000 Brussel

DANMARK

Schultz EF-publikationer
Møntergade 19
1116 København K
Tlf: (01) 14 11 95
Telecopier: (01) 32 75 11

BR DEUTSCHLAND

Bundesanzeiger Verlag
Breite Straße
Postfach 10 80 06
5000 Köln 1
Tel. (02 21) 20 29-0
Fernschreiber: ANZEIGER BONN 8 882 595
Telecopierer: 20 29 278

GREECE

G.C. Eleftheroudakis SA
International Bookstore
4 Nikis Street
105 63 Athens
Tel. 322 22 55
Telex 219410 ELEF

Sub-agent for Northern Greece:

Molho's Bookstore
The Business Bookshop
10 Tsimiski Street
Thessaloniki
Tel. 275 271
Telex 412885 LIMO

ESPAÑA

Boletín Oficial del Estado
Trafalgar 27
28010 Madrid
Tel. (91) 446 60 00

Mundi-Prensa Libros, S.A.
Castelló 37
28001 Madrid
Tel. (91) 431 33 99 (Libros)
431 32 22 (Suscripciones)
435 36 37 (Dirección)
Télex 49370-MPLI-E

FRANCE

Journal officiel
Service des publications
des Communautés européennes
26, rue Desaix
75727 Paris Cedex 15
Tél. (1) 45 78 61 39

IRELAND

Government Publications Sales Office
Sun Alliance House
Molesworth Street
Dublin 2
Tel. 71 03 09

or by post

Government Stationery Office

Publications Section
6th floor
Bishop Street
Dublin 8
Tel. 78 16 66

ITALIA

Licosa Spa
Via Lamarmora, 45
Casella postale 552
50 121 Firenze
Tel. 57 97 51
Telex 570466 LICOSA I
CCP 343 509

Subagenti:

Libreria scientifica Lucio de Biasio - AEIOU
Via Meravigli, 16
20 123 Milano
Tel. 80 76 79

Libreria Tassi
Via A. Farnese, 28
00 192 Roma
Tel. 31 05 90

Libreria giuridica
Via 12 Ottobre, 172/R
16 121 Genova
Tel. 59 56 93

GRAND-DUCHÉ DE LUXEMBOURG
et autres pays / and other countries

Office des publications officielles
des Communautés européennes
2, rue Mercier
L-2985 Luxembourg
Tél. 49 92 81
Télex PUBOF LU 1324 b
CCP 19190-81
CC bancaire BIL 8-109/6003/200

Abonnements / Subscriptions

Messageries Paul Kraus
11, rue Christophe Plantin
L-2339 Luxembourg
Tél. 49 98 888
Télex 2515
CCP 49242-63

NEDERLAND

Staatsdrukkerij- en uitgeverijbedrijf
Christoffel Plantijnstraat
Postbus 20014
2500 EA 's-Gravenhage
Tel. (070) 78 98 80 (bestellingen)

PORTUGAL

Imprensa Nacional
Casa da Moeda, E. P.
Rua D. Francisco Manuel de Melo, 5
1092 Lisboa Codex
Tel. 69 34 14
Telex 15328 INCM

Distribuidora Livros Bertrand Lda.

Grupo Bertrand, SARL
Rua das Terras dos Vales, 4-A
Apart. 37
2700 Amadora CODEX
Tel. 493 90 50 - 494 87 88
Telex 15798 BERDIS

UNITED KINGDOM

HM Stationery Office
HMSO Publications Centre
51 Nine Elms Lane
London SW8 5DR
Tel. (01) 211 56 56

Sub-agent:

Alan Armstrong & Associates Ltd
72 Park Road
London NW1 4SH
Tel. (01) 723 39 02
Telex 297635 AAALTD G

UNITED STATES OF AMERICA

European Community Information Service
2100 M Street, NW
Suite 707
Washington, DC 20037
Tel. (202) 862 9500

CANADA

Renouf Publishing Co., Ltd
61 Sparks Street
Ottawa
Ontario K1P 5R1
Tel. Toll Free 1 (800) 267 4164
Ottawa Region (613) 238 8985-6
Telex 053-4936

JAPAN

Kinokuniya Company Ltd
17-7 Shinjuku 3-Chome
Shiniuku-ku
Tokyo 160-91
Tel. (03) 354 0131

Journal Department
PO Box 55 Chitose
Tokyo 156
Tel. (03) 439 0124

NOTICE TO THE READER

All scientific and technical reports published by the Commission of the European Communities are announced in the monthly periodical **'euro abstracts'.** For subscription (1 year: BFR 3 300) please write to the address below.

Price (excluding VAT) in Luxembourg
ECU 7 BFR 300 IRL 5.50 UKL 5 USD 8.50

OFFICE FOR OFFICIAL PUBLICATIONS
OF THE EUROPEAN COMMUNITIES

L-2985 Luxembourg

ISBN 92-825-7773-2

9 789282 577738