Watson, I. & Watson, H. (1997). CAIRN: A Case-Based Document Retrieval System. In, Proceedings of the 3rd. United Kingdom Case-Based Reasoning Workshop. University of Manchester. Filer, N. & Watson, I (Eds.)

# CAIRN: a case-based document retrieval system

Ian Watson & Heather Watson
*AI-CBR, Bridgewater Building, University of Salford, Salford, M5 4WT, UK*
www.ai-cbr.org
*ian@ai-cbr.org*

**Abstract**
This paper describes a document retrieval system called CAIRN that uses case-based reasoning to retrieve documents based on natural language queries. CAIRN parses a document set using a large lexicon to automatically generate a case index to that document set. The index is used by a case-based retrieval engine to find documents. The retrieval engine is tolerant of noisy natural language queries. CAIRN also supports failure-driven learning of important concepts during its use and thus can significantly improve its retrieval accuracy over time. The limitations of this system are discussed.

## 1. Introduction

The problem of obtaining timely and relevant information from digital information sources (e.g., CD's & the Internet) is becoming daily more acute, as there is an ever increasing variety of sources and volume of information available. This problem is compounded when end-users are not familiar with search techniques. Current information retrieval (IR) systems use keyword, free-text, context-based search, hypertext structures and statistical IR techniques. These techniques are either laborious for the information publisher, since keywords and hypertext structures must be defined before publication, computationally inefficient (e.g., a free-text search of megabytes of text) or unsuitable to novice users (because of a vocabulary mismatch between how a user expresses a query and the available indexes). This paper presents initial results from the CAIRN project, which is investigating a case-based document indexing and retrieval system that can be both efficient for the content publisher and intuitive to the user.

Case-based reasoning (CBR) has recently been shown, most notably in the US electronics industry, to be a powerful and flexible retrieval technology [Kitano et al., 1992; Kitano & Shimazu, 1996], that is being increasingly used for information retrieval [Anick & Simoudis, 1993; Smail, 1993; Cunningham et al., 1995]. CBR is an improvement over some retrieval methods because it combines the friendliness of context-based searches with flexible means of indexing the information. Using case-based retrieval users can ask for information in their own words. The CBR system then asks the user for more information as required, to focus a search if too many or too few documents are returned. This dialogue between the system and the user is an intuitive way of retrieving information that lets the user remain in control.

However, most current case-based retrieval systems are manually created by a laborious process. The CAIRN project uses an *automated* indexing process, thereby greatly reducing the effort for the publisher. This approach is supported by a recent report commissioned by the Department of

Education and Employment from ICL, in which they explicitly recognise the need for AI support of the user and the librarian [Yapp, 1995, p.9].

## 2. System Requirements

At the start of the CAIRN project a set of requirements was formulated that the CAIRN tool would need to satisfy. These were divided into two categories; requirements from the users view point and requirements from the content publishers view point. Those from the users' view point were as follows:

- The tool should run on relatively low cost hardware (i.e., PCs) and be network friendly.
- It should allow users to enter free-form natural language (NL) text as a search query; e.g., a user can ask for "*advice on paints and breathing*";
- It should be tolerant of spelling mistakes and typographical errors.
- It should be reasonably quick (i.e., a second or two for retrieval).
- It should be able to infer that *thinners* and *solvent* may be synonymous.
- It should be able to ask the user questions to focus the search such as: *"Are you interested in A) solvents in confined spaces, B) safe application of spray paints C) breathing apparatus?"*.

From the publishers' view point the requirements were:

- The tool should run on relatively low cost hardware (i.e., PCs) and be network friendly.
- It should generate the case-index automatically without human intervention.
- It should be able to deal with megabytes of content.
- It should be customisable to improve retrieval accuracy.
- It should interface with existing retrieval software.
- It should have a customisable user interface.

Although there are many case-retrieval systems on the market that satisfy many of these criteria (e.g., ReMind, ReCall, and CBR3 [Watson, 1997]) generating cases by hand for all of them is very labour intensive (just as creating keywords, document summaries or hypertext links are). A recent study by the applicant showed that an experienced developer can index between 20 and 40 documents per day depending on their complexity [Watson & Abdullah, 94]. Thus, developing an index by hand for 10,000 documents could take 500 days. CAIRN automates this process and can process thousands of files per hour.

## 3. System Architecture

The CAIRN system uses the same primary software components as the Swiss Bank Corporation's Know How Project [Block & Poynter, 1996], namely Inference Corporation's CBR3 Generator and CasePoint tools. The two principle components are described below.

CAIRN reads text files (ASCII text, HTML, Microsoft Word files, and RTF files) and builds a case index to the document library. In order to understand the content of a text file CAIRN uses a lexicon containing an alphabetical list of approximately 50,000, mainly English, words identified by part of

speech and by the word's *polysemy index*[1]. In addition, content publishers can add specific technical terms, abbreviations, jargon and acronyms to the lexicon.

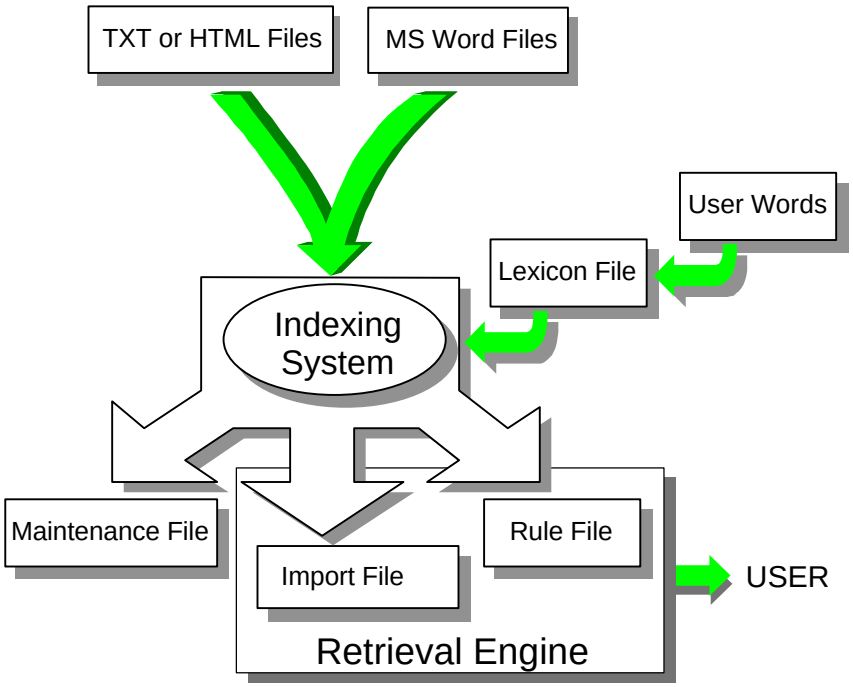


Figure 1. The CAIRN Architecture

CAIRN parses each text file using the lexicon to decide the subjects of the file (i.e., the case index). It produces a short textual summary of the file that is used in the first phase of document retrieval. It also construct a set of questions that further discriminates between files and thus focuses the search using knowledge guided retrieval. This is an important distinction from conventional IR systems which expect to receive a *well-formed* query. In CAIRN a query is refined in an iterative process of questioning and researching.

CAIRN outputs three files:

1. an indexed Import File that is in a format that can be imported into the case-based retrieval tool.
2. a Rule File that can be used by the CBR tool to ask questions and focus document retrieval.
3. a Maintenance File that will be used to maintain the case index when files in the library are added, removed or modified.

## 4. Indexing

One of the CAIRN project's collaborators is Chapman & Hall Publishers (http://www.chaphall.com/chaphall.html). They are a major multinational publishing firm with a wide range of periodicals, and technical publications. Increasingly they are using digital media to publish, including CD-ROM and the Internet. Their Electronic Publishing Division provided the

[1] The *polysemy index* of a word is a measure of the number of meanings the word can take. For example, the word *watch* is both a noun and a verb with several different meanings. This means that *watch* might be ambiguous as an indicator of document content. Conversely, the word *asbestos*, is a noun with a single meaning. Documents that contain the word *asbestos* are certainly in some way concerned with *asbestos.*

CAIRN project with data from a database of European Health and Safety Legislation. This data is currently retailed on a CD-ROM using a conventional keyword retrieval system. An sample abstract (ref. # 212880-1) is shown in Figure 2.

```
<newdocument>
<uid>212880-1</uid>
<tit>Electrical Equipment for Explosives Atmospheres (Certification) (Amendment) Regulations
(Northern Ireland) 1995</tit>
<sht>Explosive Atmospheres Regulations 1995</sht>
<lan>English</lan>
<amd>Electrical Equipment for Explosive Atmospheres (Certification) (Amendment) Regulations
(Northern Ireland) 1991 SR 1991 No. 339<xuid>213973-3</xuid></amd>
<sub>Equipment, Workplaces and Operations</sub>
<sub>Fire, Explosions</sub>
<sub>Industries</sub>
<des>Approval</des>
<des>Electrical Equipment</des>
<des>Explosive Atmospheres</des>
<des>Mining Industries</des>
<rel>European Communities Act 1972 <xuid>240390-7</xuid></rel>
<non>Council Directive 79/196/EEC concerning electrical equipment for use in potentially
explosive atmospheres employing certain types of protection<xuid>241784-6</xuid></non>
<abs>These Regulations amend the Electrical Equipment for Explosive Atmospheres (Certification)
Regulations (Northern Ireland) 1990. Regulation 3(3) substitutes a new Regulation 12 which now
states that a manufacturer of electrical equipment who applies for a certificate of conformity will be
able to have that equipment certified by reference to the harmonised standards in the unamended
regulations. Where an application for a certificate of conformity or review is made before the coming
into force of these regulations such application shall continue to be dealt with under the unamended
regulations. Certificates of conformity will not be issued in respect of electrical equipment to which
the Framework Directive applies after 29 February 1996 or to which the Gassy Mines Directive
applies after 31 December 1996. Certificates of conformity issued on or before these dates shall be
regarded as in force as regards the use of a distinctive Community Mark until 30 June 2003.</abs>
```

Figure 2. Sample Abstract from Health & Safety Data Set

As you can see the text is quite noisy, with a lot of formatting tags in it. The most important fields in the sample are the Subject <sub> and Description <des>. These were defined manually by Chapman and Hall and include the following terms: *Equipment, Workplaces, Operations, Fire, Explosions, Industries, Approval, Electrical Equipment, Explosive Atmospheres, Mining Industries*. These are the keywords that a user would search on to retrieve this piece of legislation. CAIRN automatically produced the case shown in Figure 3 to index this file.

The case is divided into four main areas. The *Title* and Description fields are what are matched against the user's NL-query. The questions are used to confirm or reject this case (i.e., to discriminate this case from other similar cases). The attachments will open the appropriate document (i.e., the file 212880-1.TXT).

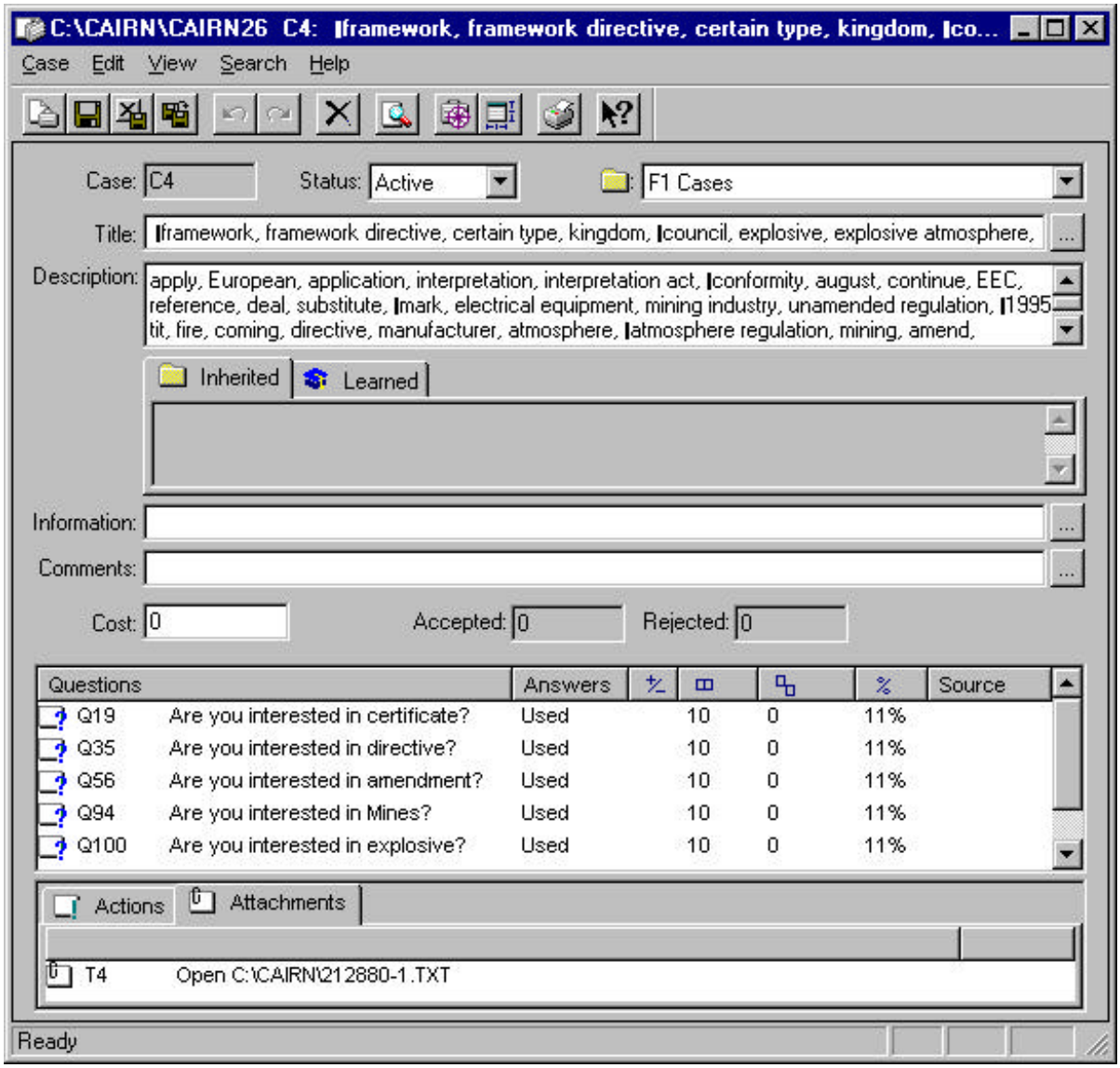


Figure 3. An Automatically Generated Case

One way of assessing the utility of the automatic case index generation is to determine if CAIRN accurately identifies the same keywords as the publishers did manually. Thus, one would expect CAIRN's title and description to contain the publisher's keywords. A comparison is shown in Table 1.

Of these terms only *workplaces* is not present. The singular of explosions, atmospheres and industries are all present, and since CAIRN is not adversely effected by the differences between singular and plural terms these can be read as the same. However, more strikingly is the fact that of the five confirming questions that CAIRN generated two are about *mines* and *explosives.* This confirms that CAIRN has correctly inferred that this piece of legislation is a directive about mines and explosions. A study of 5000 cases showed that in 95% of cases, 90% or more of the publisher defined keywords were contained in the automatically generated case title, description or questions. This confirms that CAIRN is accurately indexing the text files.

| Publisher's keywords | CAIRN's title & description keywords |
|---|---|
| Equipment, Workplaces, Operations, Fire, Explosions, Industries, Approval, Electrical Equipment, Explosive Atmospheres, Mining Industries | framework, framework directive, certain type, kingdom, council, explosive, explosive atmosphere, non council, firedamp, commission, commission directive, explosion, explosion sub, february, certificate, certification, date, standard, potential, apply, European, application, interpretation, interpretation act, conformity, august, continue, EEC, reference, deal, substitute, mark, electrical equipment, mining industry, unamended regulation, 1995, fire, coming, directive, manufacturer, atmosphere, atmosphere regulation, mining, amend |

Table 1. Manual vs. Automatically Generated Keywords

## 5. Retrieval

Once the case-index has been generated Inference's CBR3 CasePoint retrieval engine is used. A user can enter a NL-query into the description field and CAIRN will search the case-index for cases whose titles and descriptions are similar to the NL-query. Using the same text file as our target (i.e., 212880-1.TXT concerning explosives and mines) a user could type in the simple query "*explosives*". This is shown in Figure 4.

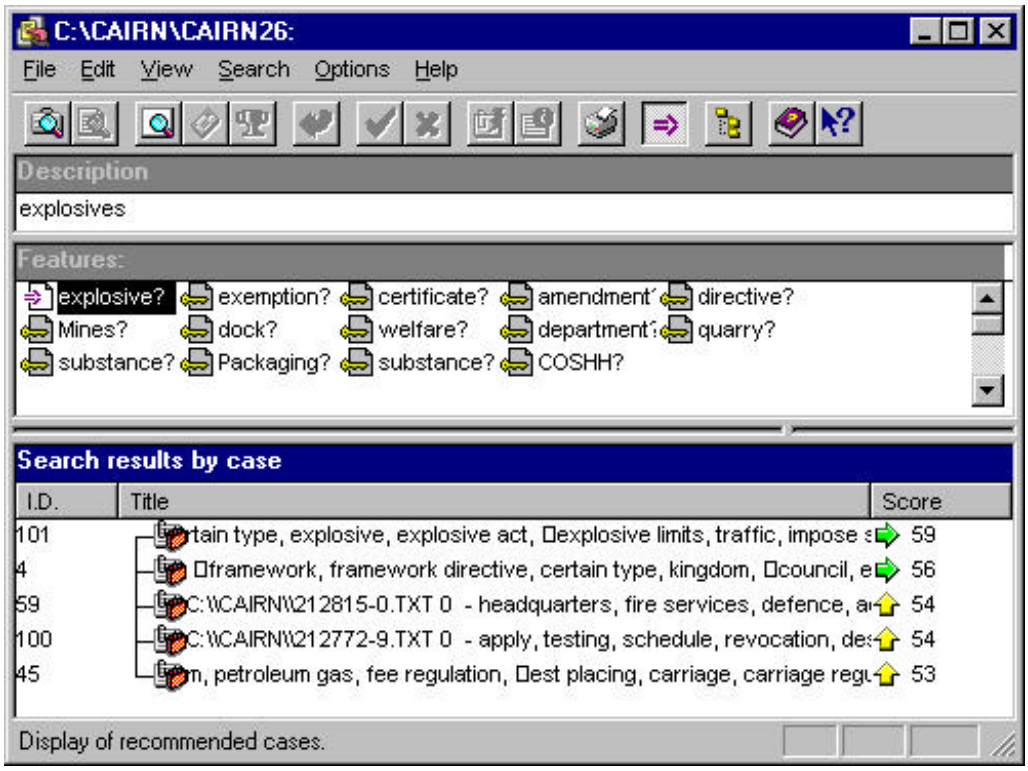


Figure 4. Retrieval Using a Simple Query

At the bottom of the CAIRN screen is a list of cases ranked in order of similarity. In the middle are a set of features that help discriminate between the cases. The text we are looking for is indexed by case id. 4. If the user clicks the mouse on *Mines?* in the feature panel the similarity score for case id. 4 increases from 58 to 67. This is shown in Figure 5.

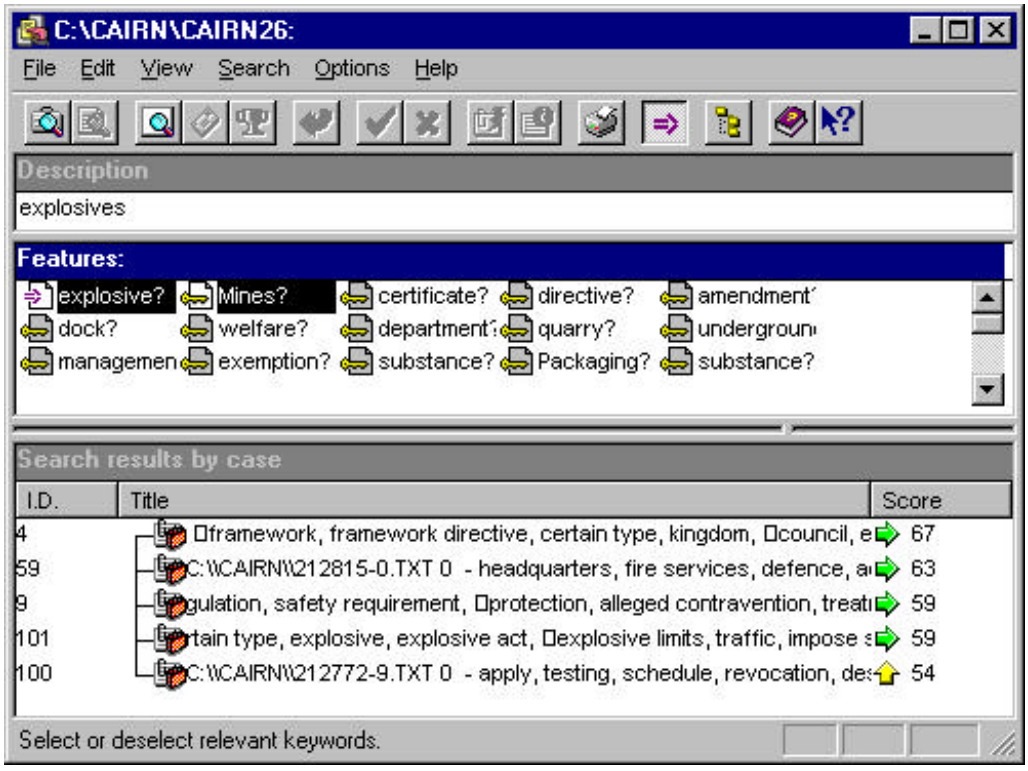


Figure 5. Interactive Retrieval

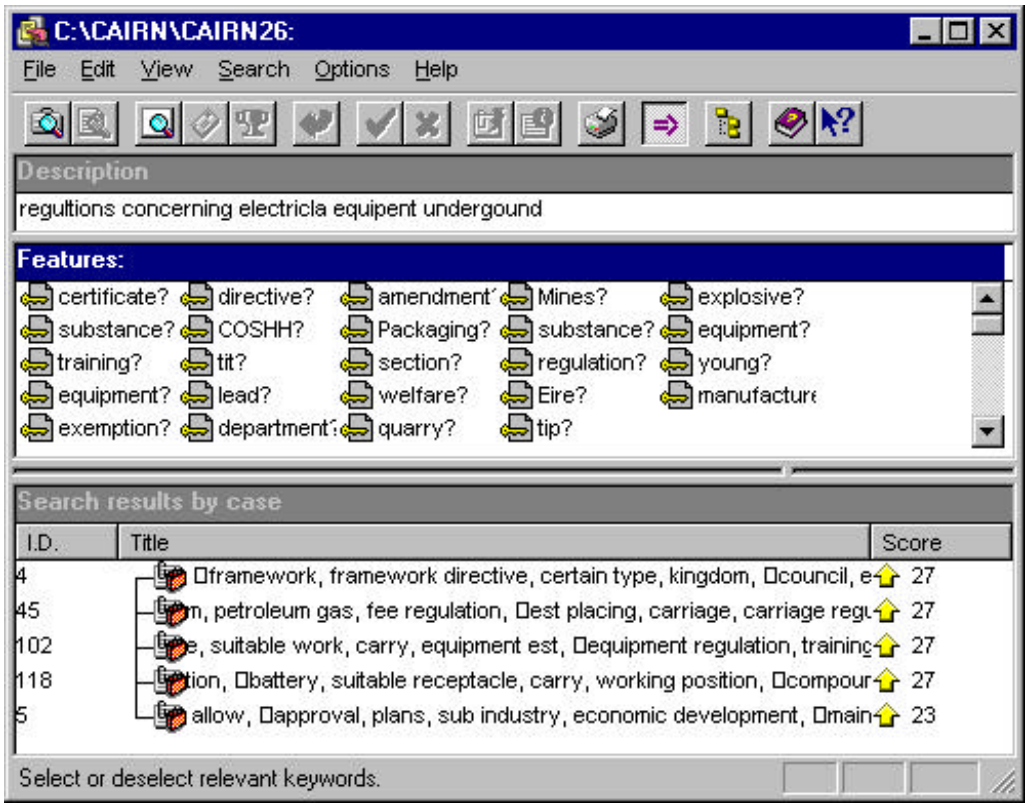


Figure 6. Retrieval Using a Noisy Query

Thus, CAIRN has correctly found the target document. Let us now try a more difficult and noisy query such as "*regultions concerning electricla equipent undergound*" (the spelling mistake are intentional in this query). Given this query CAIRN performs well as is shown below.

CAIRN has successfully ranked case id. 4 at the top of the similar cases despite the incorrect spelling of many terms in the query. Finally, if CAIRN is given a precise and accurate query such as might be provided by someone who was familiar with the data source (e.g., "*directives about electrical equipment in explosive atmospheres in mines*"). It performs very well as shown in Figure 7.

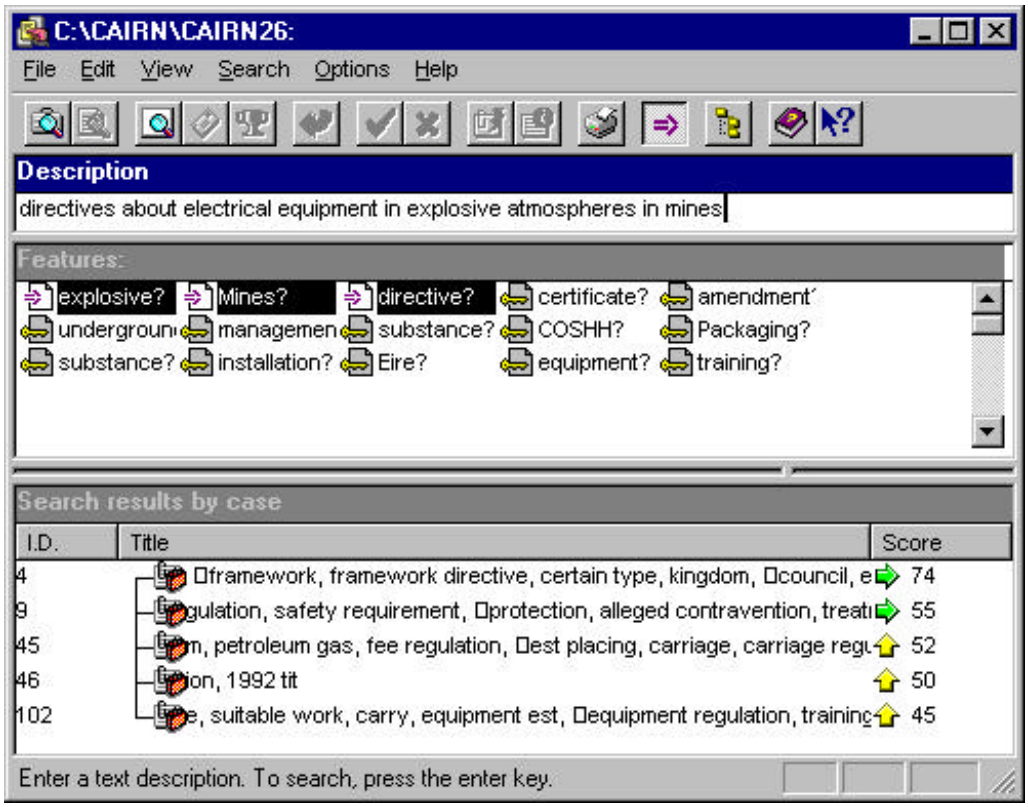


Figure 7. Retrieval Using a Complex Precise Query

## 6. Failure-driven learning

CAIRN has one other useful feature, that of failure-driven learning. A document retrieval system will probably never perform at 100% accuracy, thus it is useful if a retrieval system can learn from its failures. This is a characteristic of CBR, which is called failure-driven learning [Schank, 1982; Leake 1996]. If CAIRN fails to correctly retrieve a document, the user can resort to using another search technique (e.g., an exhaustive word search of the document set, such as that provided by the Windows 95™ "Find" utility or by using other search engines such as Verity). If a document is successfully retrieved (i.e., one that matches the initial NL-query), CAIRN records the concepts in the original NL-query and adds these to the "Learned" field of the case index (shown in Figure 10). These *learned concepts* can then be used by CAIRN as part of the case index. Thus, over time, the performance of the retrieval engine will improve.

Failure-driven learning was tested on a document set of abstracts on aerodynamics. The set of 1400 abstracts was obtained from the IDOMENEUS technology transfer server in the Department of Computing Science at the University of Glasgow (http://www.dcs.gla.ac.uk/idom/). The Cranfield test collection, comes with standard NL-queries and list of documents that are relevant to each query. Thus it was possible to compare CAIRN's performance against that of the test set. These

abstracts are of a highly technical nature with many terms that are not in CAIRN's lexicon. Moreover, many of the documents are similar making it difficult for a search engine to differentiate between them.

> periodic temperature distribution in a two-layer composite slab.
> w. f. campbell
> national aeronautical establishment, ottawa, ont., canada
> in a recent contribution to the reader's forum, under the above title, stonecypher outlined a method for finding the periodic temperature distribution in a two-layer composite slab, one exposed surface of the slab being insulated and the other subject to a sinusoidal temperature variation. perfect thermal contact between the two layers, and constant thermal properties were assumed. two years ago i drew attention in these pages to a method for determining the transient temperature in such a two-layer slab resulting from a triangular heat-input pulse . i should like to point out that this same method also is applicable to the case where one external face is given a sinusoidal temperature variation with time. the method is based on the analogy between one-dimensional heat flow and the flow of an electric current in a simple transmission line having only series resistance and parallel capacitance.

Figure 8. Sample Abstract from the Cranfield Data Set

Consequently, at first, CAIRN's performance was poor (an average 20% accuracy rate). Over successive generations (i.e., running the queries and manually retrieving the correct documents, if necessary so CAIRN could learn new concepts) CAIRN's performance improved to almost a 60% accuracy. These results are shown in Figure 9. The slow start to concept learning is explainable. As CAIRN learns new concepts, it initially does not give them much credence until it has seen the same concept appearing several times in the NL-query. This is a user definable threshold that was set to 3 (i.e., don't give a new concept much importance in retrieval until the new concept has occurred at least three times in unsuccessful searches). It is also predictable that concept learning would plateau since a limited number of NL-queries were being used. It was perhaps disappointing that the plateau was not higher than 60%.

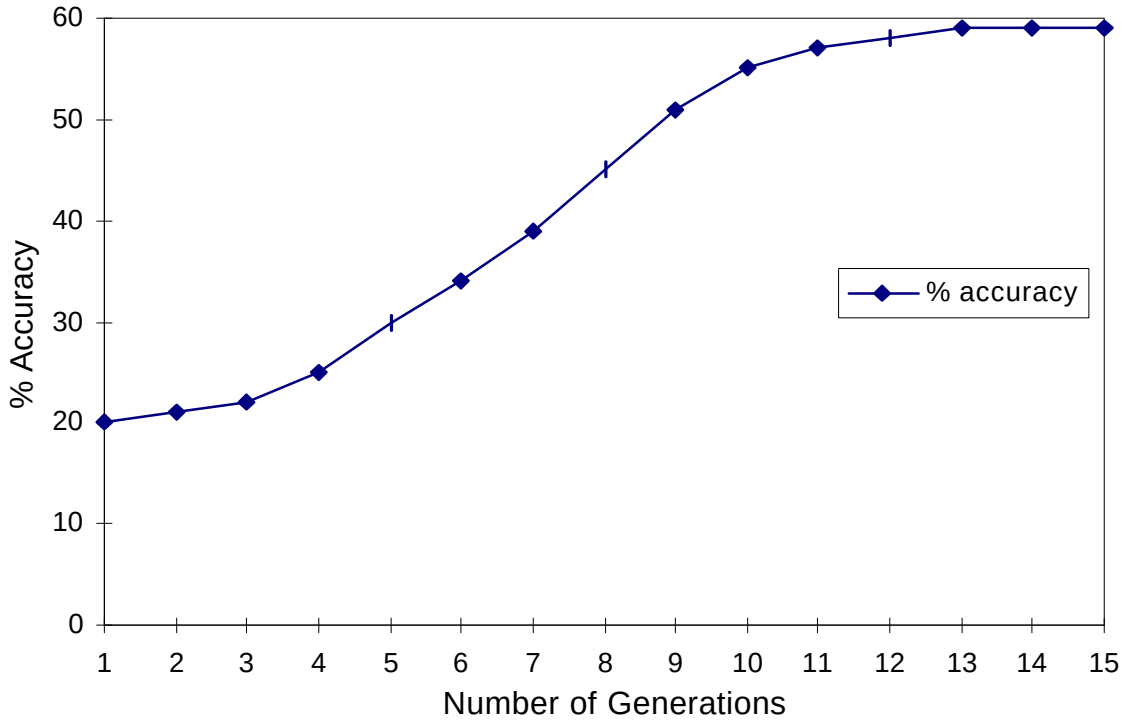


Figure 9. Graph of Concept Learning Indicated by % Accuracy

Thus, it is clear that to achieve a higher success rate it would be necessary to add many technical terms to CAIRN's lexicon at the outset (i.e., before the automatic generation of the case index).

Concept learning certainly helps but for this data set it was not enough. This is the next phase of the project. Figure 10 shows the concepts that CAIRN learned from the queries for the abstract shown in Figure 8.

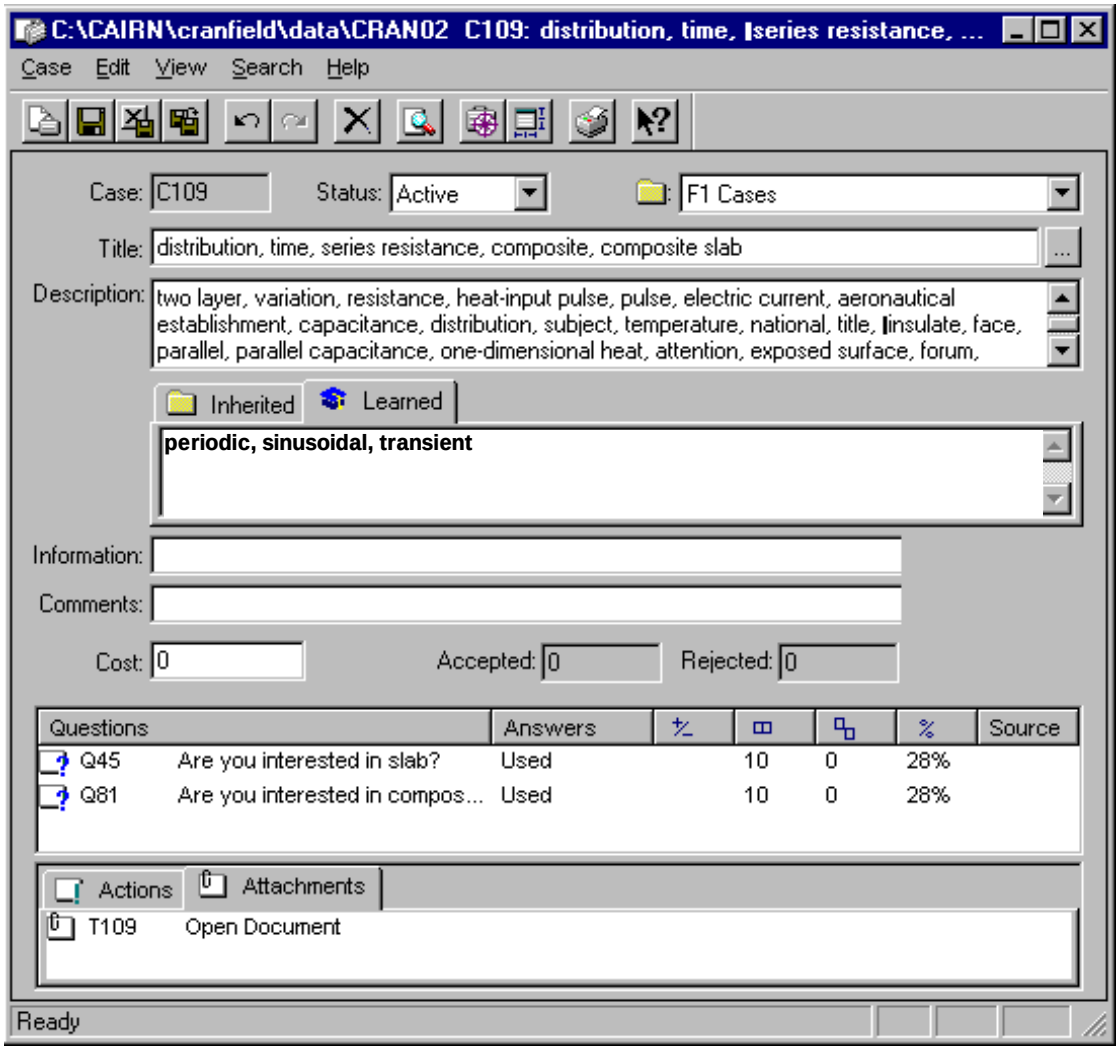


Figure 10. The Case Editor Showing Learned Concepts

## 7. Conclusions

Our initial evaluation of CAIRN has shown that it immediately performs well on document sets such as the Health & Safety legislation data set provided by Chapman & Hall Publishers, despite the presence of noise in the documents. Its tolerance of noise is not surprising, since when CAIRN parses a document it ignores anything it does not recognise as a word in its lexicon. More problems were met on highly technical data sets such as the Cranfield aerodynamics data set. Initially, CAIRN's performance was poor. Even after allowing a significant time for failure-driven learning its accuracy was only at best tolerable. This clearly indicates that for highly technical documents CAIRN's lexicon needs to be added to, for its performance to be acceptable. Whilst it can learn new index concepts and improve its performance through failure-driven learning, it is doubtful if users would continue using a search engine that performed so poorly at first. Another significant limitation encountered is the size of the data set that CAIRN can feasibly index. Although we have not performed a rigorous study, our empirical evidence would suggest that as the data set

approaches 5000 documents the time taken to index the documents becomes unacceptable. When CAIRN produces the case to reference the document it seeks to differentiate the new case from others in its index. Thus, as the size of the case index increases the time taken to complete this task would appear to increase at least geometrically. However, despite these limitations the approach demonstrated here does show good results with appropriate data sets and has met all of our initial requirements, except the ability to handle many megabytes of data. We shall continue to experiment with case-based information retrieval.

## 8. Acknowledgements

This work was funded by the UK Engineering & Physical Sciences Research Council, Grant Number GR/L16330.

## 9. References

Anick, P., & Simoudis, E. (Eds.) (1993). *Case-Based Reasoning & Information Retrieval: Exploring Opportunities for Technology Sharing*. AAAI Spring Symposia SS-93-07

Block, F., & Poynter, L. (1996). The Swiss Bank Corporation Know How Project. In, *Applications & Innovations in Expert Systems IV*. Macintosh, A., & Cooper, C. (Eds.), SGES Publications, UK.

Cunningham, P., Bonzano, A., & Smyth, B. (1995). *An Incremental Case Retrieval Mechanism for Diagnosis*. Technical Report, TCD-CS-95-01. Dept. of Computer Science, Trinity College, Dublin.

Kitano, H., Shibata, A., Shimazu, H., Kajihara, J., & Sato, A. (1992). Building large-scale and corporate wide case-based systems. In, *Proceedings of AAAI-92.* Cambridge, MA: AAAI Press/MIT Press.

Kitano, H., & Shimazu, H. (1996). The Experience Sharing Architecture: A Case Study in Corporate-Wide Case-Based Software Quality Control. In, *Case-Based Reasoning: Experiences, Lessons, & Future Directions*, Leake, D.B. (Ed.). AAAI Press / The MIT Press, Menlo Park, Calif., US.

Leake, D.B. (Ed.). (1996). *Case-Based Reasoning: Experiences, Lessons, & Future Directions*. AAAI Press / The MIT Press. ISBN 0-262—62110-X

Schank, R. (Ed.) (1982). *Dynamic Memory: A Theory of Learning in Computers and People*. New York: Cambridge University Press.

Smail, M. (1993). Case-Based Information Retrieval. In, *Proceedings of EWCBR'93.* Richter, M.M., et al. (Eds.), Springer-Verlag.

Watson, I. (1997). *Applying Case-Based Reasoning: Technology for Enterprise Systems*. Morgan Kaufmann Publishers Inc., Menlo Park, San Francisco, CA.

Watson, I., & Abdullah, S. (1994). Developing Case-Based Reasoning Systems: A Case Study in Diagnosing Building Defects. In, *Proceedings of the IEE Colloquium on Case-Based Reasoning: Prospects for Applications*, Digest No: 1994/057, pp.l/1-1/3.

Yapp, C. (1995). *The Training Superhighway: Re-engineering training to support Life Long Learning*, UK Department of Education & Employment, September, 1995.

Watson, I. & Watson, H. (1997). CAIRN: A Case-Based Document Retrieval System. In, Proceedings of the 3rd. United Kingdom Case-Based Reasoning Workshop. University of Manchester. Filer, N. & Watson, I (Eds.)

Information on all aspects of case-based reasoning can be found at www.ai-cbr.org