CASE-BASED REASONING MODEL APPLIED TO INFORMATION RETRIEVAL

Carlos Ramirez and Roger Cooley

Computing Laboratory
University of Kent at Canterbury
Canterbury, Kent.
{cr10, rec}@ukc.ac.uk

ABSTRACT

This paper is about work in progress on the application of Case Based Reasoning (CBR) to problems of Information Retrieval (IR). Many approaches to IR have already been developed. They involve techniques such as keyword searches and the use of boolean queries which are applied to databases or indexes. Other methods are based on probabilistic or statistical analysis of occurrences of words or phrases, some on structured queries, and yet others work with vector space models or semantic models. Knowledge-based approaches are not uncommon. For example, expert intermediary systems may be used to formulating and evaluating queries. However, there is little experimental evidence to demonstrate the effectiveness of knowledge-based techniques and they remain a significant challenge for research [Croft, 1993].

The main idea under Case-Based Reasoning (CBR) is to store experience as cases, to store problem-solving processes as instances of cases. When a new problem is encountered, the system uses its memory of relevant past cases to interpret or solve the problem. Adaptation of cases plays an important role in this process. Thus, the more cases the system stores the better it performs. CBR allows exceptional cases to be stored and used in problem solving, makes learning from cases possible, and provides more convincing explanations, based upon the stored cases. Moreover, cases can be used to establish general rules. Due to this characteristics, CBR is specially useful in areas that (a) are particularly difficult to formalise (b) planning and explanations are required (c) persuasion is important, or (d) hypothetical scenarios are needed.

Learning and adaptation have long been viewed as crucial part of an IR system [Salton, 1983], and this alone suggests that CBR might prove to be a valuable approach to this area. In this paper, a CBR model is proposed as an alternative approach to to IR. The possible relationships of a CBR system to the various facets of current human and automated systems are considered, with particular attention being paid to classificatory schemes and keyword searches. However, the focus of the paper is on issues of representation and adaptation. Close attention is paid to what exactly should constitute a case; and a system is described which starts from a problem description, ie a need for information, and a context, which involves both intellectual and institutional features. Learning takes place by adding new cases and modifying old cases in the system's case-base.

1. Overview

This paper is about work in progress on the application of Case Based Reasoning (CBR) to problems of Information Retrieval (IR). Many approaches to IR have already been developed. Traditional models involve techniques such as keyword searches and the use of boolean queries which are applied to databases or indexes. Other methods are based on probabilistic or statistical analysis of occurrences of words or phrases, some on structured queries, and yet others work with vector space models or semantic models. Knowledge-based approaches are not uncommon. For example, expert intermediary systems may be used to formulate and evaluate queries. However, there is little experimental evidence to demonstrate the effectiveness of knowledge-based techniques and they remain a significant challenge for research [Croft, 1993].

Adaptation and Learning have long been viewed as crucial parts of an IR system [Salton, 1983], and this alone suggests that CBR might prove to be a valuable approach to this area. In this paper, a CBR model is proposed as an alternative approach for IR. Close attention is paid to what exactly should constitute a case; and a system is described which starts from a problem description, i.e. a need for information, and a context, which involves both intellectual and institutional features.

2. Problem Description

Information Retrieval deals basically with the issue of finding information relevant to users' requests. We intend to focus our attention on collections of abstracts (which we will call "documents"). Traditional IR systems have tackled this issue using models containing three fundamental processes, which we can generalise from Croft's description [Croft, 1993]:

1. Representing the contents of the documents.
2. Representing the users' information requests. That is, formulating a query.
3. Comparing these representations to decide which document should be retrieved.

The representation of a documents is usually done by associating a set of attributes with the document, whose values are assumed to describe the contents of the document. In these circumstances, the retrieval of particular documents depends on the similarity between the attributes of the documents and those used in the formulation of a user's request. Similarity is usually assessed by trying to match attribute values and then looking for a sufficient degree of coincidence between the sets of attributes attached to queries and documents. The selection of attributes is a rather difficult issue, since it is clear that representing the contents of a complete document in a set of attributes results in a loss of precision in terms of the contents of the document. Another important issue is related to the normalisation of attributes, a problem arises due to the possibility of representing attributes with more than one term or value (e.g. by the use of synonyms or related terms). This may be due to users who are not familiar with document indexing structures or who are not familiar with query composition, (e.g. the use of boolean operators, the manipulation of sets of elements, etc.). Thus, a very useful complementary approach in IR systems is the expansion of queries through the use of a lexicon and additional mechanisms (e.g., relevance feedback). This, improves the overall recall/precision of a system; however, there is significant scope for further improvement [Harman, 1992; Hains, 1993].

3. Mapping Information Requests to Documents

As we saw above, a fundamental problem confronted by ordinary users of IR systems is the formulation of appropriate requests. A request formulation must include the necessary features to acquire relevant information from a database. This is difficult, and it is uncommon to retrieve a good number of relevant documents on the first attempt, thus a process of trial and error is necessary to improve the query formulation.

We have developed a CBR model to tackle IR problems, which uses an improved representation of the users' needs and also works on query refinement. This model uses an iterative mapping of the User Information Need Space (represented in a request by a Query plus a Context), onto the Information Retrieval Space (represented as a database of documents) as shown in Figure 1. The refinement process is carried out on each retrieval trial by either transforming the user's request or by transforming the corresponding case instance; thus generating the next case instance.

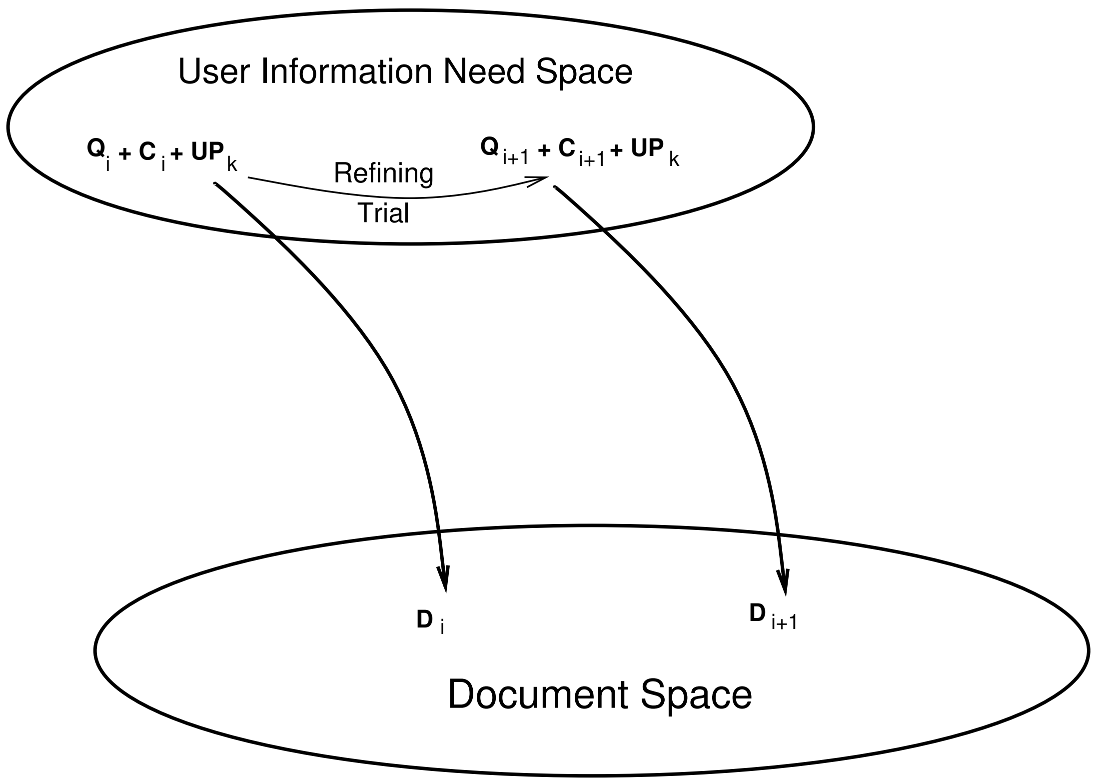

Fig. 1. Mapping Information Requests to Documents

4. Context and Queries

A user's query is represented in the CBR model as a series of query terms in the same way as it might be in a traditional system. Additionally, the model includes a representation of information about the context in which the query arises. In a university setting, for example, the context information might include the fact that a query is connected with a particular academic course or even a particular assessment exercise. Such context information is routinely used by librarians but it is not incorporated in IR systems.

5. Case Representation

Our goal is to respond to users with lists of references to documents relevant to their requests. The two main advantages of using CBR to tackle the problem are:

(a) The ability to use and build on past experiences of users' information requests.
(b) The ability to incorporate context information.

In order to do this, we need to be able to represent the necessary information in suitable storage structures. Thus, we defined two different case bases to conveniently store and manipulate the information with maximum flexibility: The first case base, named *Query-Case Base* (QCB), is used to store queries and their different transformations. A *Query* (Q) is the main representation of the request, is the users explicit description of their information needs, expressed by search terms. The second case base, named *Context-Case Base* (CCB), is used to store *Context* (C), the context in which every case is embedded. For our particular problem, context is a set of attributes describing features of the request that are not included in the query. We'll also have a support file: the *User's Profile* (UP) containing information, which will act as adjustments features, as described below.

A *Query-Case* is build up by a set of attributes representing the explicit features of users' information need. The features to be described are the various stages in the formulation of a query. They may be up to four stages which are referred to as Initial, First, Second and Final Refined Features. The Initial features are search terms taken directly from the Query Description provided by the user. The First, Second and Final features are refinements generated by the system. Then we have Case References to similar cases. Finally, the attribute Results used to store references to documents.

A *Context-Case* is a set of attributes representing not-explicit features of the request that are not included in the Query-Case, but which are relevant for query refinement and for adjustment operations. Those features are represented in the following attributes: Related Subjects, Related Topics, Objectives of the Information, Particular Application or Use.

A *Users'-Profile* contains information that isn't updated on every request, but only when users' relevant activities change. The attributes of this file contain features about users' Professions or Activities, about their Current Areas of Work and Related Areas or Subjects, on their Current Areas of Specialisation and Topics Related to them and Particular Work Objectives for the information requested. So, the attributes of this file can be used to substitute context features when they are not present in the CCB or when new search terms still need to be generated after exhausting the features in the QCB and the CCB.

Thus, a case is defined in the following expression:

$$CASE = \{Q_{i,}\ C_{i,}\ UP\}, \quad i = 1, \ldots, 4$$

In [Harman, 1992] it is suggested that $i<4$, since there seems to be no significant improvement after 3 trials and we don't want to store unnecessary instances of a case, because of case retrieval efficiency.

This information is used as follows. The CBR system first attempts to satisfy a user's request by accessing the QCB using only the query terms provided to find a similar request. If this is successful, the user can be given references to documents without searching the collection of documents. When several similar cases are retrieved, the context information is used to determine the most likely case. Failure to find a similar case in the QCB, however, will initially lead to a refinement of the query which exploits the context information, but which may eventually lead to the system prompting the user for additional search terms and directly searching the document database. This process of query refinement results in the construction of an additional case. This is how the system's performance improves with use.

6. Acknowledgements

This work is supported by The Consejo Nacional de Ciencia y Tecnologia (CONACYT) and The Tecnologico de Monterrey,Campus Queretaro (ITESM), Mexico.

7. REFERENCES

[Croft, 1993] Croft W.B., Knowledge-Based and Statistical Approaches to Text Retrieval, IEEE Expert, April 1993, p. 8-11.

[Harman, 1992] Harman, D. Relevance Feedback Revisited. Proceedings of the 15th International Conference on Research and Development in IR, ACM SIGIR 1992.

[Hains et al., 1993] Hains, D., Croft W.B., Relevance Feedback and Inference Networks, Proceedings of the 16th International Conference on Research and Development in IR, ACM SIGIR 1993.

[Salton, 1983] Salton, G., McGill, M.J. Introduction to Modern Information Retrieval, McGraw Hill, New York, 1983.