Evaluation of an Ontology-Based Information Retrieval Tool

Stuart Aitken[1] **and Sandy Reid**[2]

Abstract. This paper evaluates the use of an explicit domain ontology in an information retrieval tool. The evaluation compares the performance of ontology-enhanced retrieval with keyword retrieval for a fixed set of queries across several data sets. The robustness of the IR approach is assessed by comparing the performance of the tool on the original data set with that on previously unseen data.

1 Introduction

Interest in the evaluation of knowledge-based systems (KBS), knowledge modelling and knowledge acquisition (KA) tools and techniques is growing due to the increased maturity of the field, and to pressure from funding bodies and commerce for measurable costs and benefits of knowledge technologies.

KBS researchers have studied the problems of assessing knowledge reuse, the effectiveness of KA tools, and the adequacy of acquired knowledge. Quantitative studies are relatively rare, the exceptions include [2, 8, 3], and the norm is a report of the qualitative benefit of e.g. the use of an existing ontology, or the use of a KA tool. The applicability of software engineering (SE) approaches to KBS evaluation has been noted previously [6] and such techniques been successfully applied [4, 7].

This paper presents a quantitative evaluation of the CB-IR information retrieval (IR) tool developed for a UK engineering company, BAE Systems. The tool uses both keyword and ontology-based word matching to recall records which mostly consist of free text. The evaluation addresses system performance. System design and knowledge acquisition and reuse are not addressed.

We begin by describing CB-IR, then present the evaluation methodology in Section 3. The results are given in Section 3 and finally some conclusions are drawn.

2 Application and the CB-IR System

The application is the analysis of defect reports which are filed for automated test equipment (ATE) systems. ATEs are complex assemblies of signal analysers, power supplies and switching units that are used to test high-integrity radar and missile systems. Records of all calibration, repair and replacement of ATE components are centrally maintained. While the records have a standard format, the most important information is in a *Remarks* field, which consists of brief notes of what was found to be wrong or what work was carried out. A typical example of a *Remarks* field is:

```
fails self test block urv4. drawer removed
and refitted
```

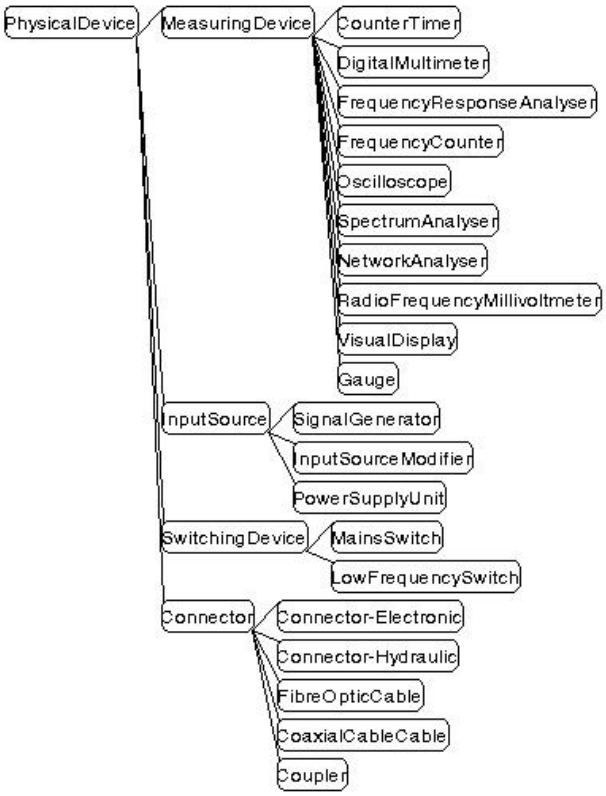

Figure 1. Excerpt of the domain ontology

The notes are often ungrammatical, and typically use abbreviations to refer to components. For example, *power supply*, *psu* and *ps* all refer to power supply units. In addition, ATEs may have several power supply units and different models have different PSU components - the records do not necessarily contain clear references to unit types. The purpose of the CB-IR tool[3] is to enable the reports to be analysed automatically by providing a query-driven recall mechanism.

CB-IR makes use of an ontology of the domain which consists of a taxonomy of domain concepts (an *isa* hierarchy) and a *part-of* relation. The lexical terms which are used to identify concepts in free text are also part of the ontology. The ontology was acquired by a

[1] AIAI,University of Edinburgh, Edinburgh EH1 1HN.
[2] BAE Systems, South Gyle Industrial Estate, Edinburgh EH12 9EA

[3] As we consider records to be **cases**, the tool was developed in the case-based reasoning paradigm.

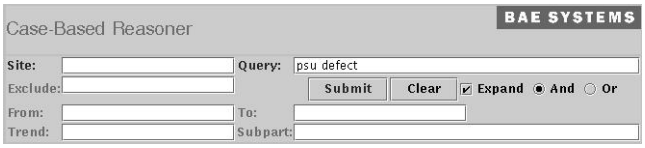

Figure 2. Query interface

manual analysis of records and documentation, and refined by consultation with domain experts. The retrieval tool uses both keyword and ontology-based word matching to match queries with records, both of which consist of free text.

A small part of the PhysicalDevice hierarchy of the ontology is shown in Figure 1. Devices such as CounterTimer have several more specific subclasses, but these are not indicated for the sake of clarity. The ontology contains 244 concepts (classes), 264 subclass relations, 104 part-of relations, and 256 concept to term relations.

The lexical terms which are associated with concepts are used to extract ontology concepts from the text fields of records. The same procedure is applied to the users' query. This means that the mapping from query to ontology can be hidden from the user, i.e. the user is free to enter free text and need have no knowledge of the ontology. Query-to-record matching is based on matching concepts, possibly using the concept hierarchy if that option has been selected.

After submitting a query, the user is presented with an ordered list of the most highly ranked matching cases. More details of each case can be viewed by a hyper-link style navigation. Queries can be refined by adding more terms and using the *and* connective, or can be widened by using the *or* connective. Alternatively, the *expand* option can used to select a concept matching procedure which uses the class hierarchy. In this case, subconcepts of query concepts which occur in cases are considered as matching the query. The logic of *and/or* and *not* as applied to query terms can be consistently combined with *expand*. This is an important point as we wish to retain the intuitive semantics of queries, i.e. concept occurs in/does not occur in.

The query interface is shown in Figure 2. *Query*, *Exclude*, *From* and *To* parameters have the usual meaning. *Site* allows the search to be restricted to a specific type of ATE, a common user requirement. *Subpart* confines the search to subcomponents of named part using the part-of model. The *Expand* checkbox and the *Subpart* parameter are the only additional elements of the interface that the user need understand in order to benefit from ontology-enhanced matching.

CB-IR also provides an ontology editor and a specialised form interface for entering records, in addition to a query interface. This enables the ontology to be extended, allowing new equipment types, or new domains to be entered. Direct manipulation techniques are combined with view selection, being the selection of a particular relation, to assist (and constrain) the user. The user is assumed to be a domain expert, and not necessarily a knowledge engineer.

3 Evaluation Method

Following the recommendations by Cohen, Menzies and others [1, 5, 6], we shall present a number of hypotheses which we wish to test. Measurements which will support the hypotheses will be defined, and the data collection process described. Our approach is strongly influenced by the application of the Goal-Question-Metric (GQM) technique by Nick, Althoff and Tautz [7] to organisational memories. We analyse the information retrieved by the system, from the viewpoint of the user, in the context of the analysis of ATE incident reports. We shall identify quality factors which support the evaluation of technical utility. However, we do not have the resources to consult the stakeholders extensively as would be recommended in a full application of GQM.

The main goal of the evaluation is to assess the technical utility of CB-IR. This goal has two constituents: assessment of the system's performance in absolute terms, and with respect to competing techniques, and assessment of the adequacy of the knowledge represented in the system regarding its impact on system performance.

The main competing technique is simply to recall cases where query words occur. Recall and precision can be measured for both ontology and keyword techniques. This allows absolute and relative measures of performance to be calculated using standard measures.

Assessment of the adequacy of the ontology can be made by measuring recall and precision for queries which explicitly use the more abstract ontology classes. This is an evaluation of the ontology from the perspective of its use in an information retrieval application. An assessment of the design of the ontology could be made as described in [3].

The robustness of concept extraction is a factor in the assessment of the ontology, hence there is a need to assess adequacy and comparative performance on both the data set for which the system was constructed (and can be expected to perform well on) and on new data sets.

The measurement of recall and precision requires reference to a human assessment of the relevance of a case to a query. Due to the effort involved in such an analysis, the number of queries that can be assessed is limited. Making such assessments on a large scale is itself an error prone process and we shall make efforts to counteract this. We also know a-priori that there is a range of standardisation in the words used in the data sets which we have analysed to date. For example, devices may have many synonyms, but the outcome of cases is stated in 3 or 4 standard terms. Therefore we must consider possible biases in the choice of queries used in the competing techniques test. The queries used to assess adequacy should not be too abstract, e.g. mention the concept *Thing*, or too specific, they should test the intermediate categorisations. The choice of data set and query would be classified as variation factors in the GQM approach. The final choice of queries is as follows:

Query No.	Query Description
Q1	Does the case mention the concept *defect*.
Q2	Does the case refer to a *Power Supply Unit*.
Q3	Does the case refer to a *Digital MultiMeter*.
Q4	Does the case refer to any instance of a *Measuring Device*.
Q5	Does the case refer to any instance of a *Connector*.

Table 1. Test Queries

Q1-Q3 will be used in the comparative assessment. Q1 tests performance on a term which we know to have a regular usage, while Q2 and Q3 are queries about device types which have a less predictable usage in the record set.

Q4 and Q5 will be used in the adequacy assessment. These queries test the categorisation of devices by function, and the categorisation of ancillary components respectively.

The follow specific hypothesis are made for comparative perfor-

mance:
H1. recall and precision are greater for ontology-based matching than for keyword-based matching on the original data set
for adequacy:
H2. recall and precision are greater than 90% for ontology-based matching on the original data set
for robustness:
H3. recall and precision are greater for ontology-based matching than for keyword based matching on the new data sets
H4. recall and precision are greater than 80% for ontology-based matching on the new data sets

Keyword and ontology-based queries will be run on three data sets, in support of hypotheses H1-H4. The matrix of query against data set is given below. DS1 is one of data sets used in the original KA phase of development, DS2 and DS3 are additional data sets, the table entries are the hypotheses that will be supported by collecting this data. Queries can be run as keyword (KWD) or ontology based matching (ONT).

Query No.	DS1	DS2	DS3
Q1 ONT	H1,H2	H3,H4	H3,H4
Q1 KWD	H1	H3	H3
Q2 ONT	H1,H2	H3,H4	H3,H4
Q2 KWD	H1	H3	H3
Q3 ONT	H1,H2	H3,H4	H3,H4
Q3 KWD	H1	H3	H3
Q4 ONT	H2	H4	H4
Q4 KWD	-	-	-
Q5 ONT	H2	H4	H4
Q5 KWD	-	-	-

Table 2. Query/Data Set/Hypothesis matrix

The keywords used are: Q1 `defect`, Q2 `psu`, Q3 `dmm` as these are the most commonly used abbreviations. Any case sensitivity in the data will be removed prior to testing.

The final element of the evaluation is to determine how the baseline assessment will be carried out. We adopt a simple approach where the human reviewer is presented with a questionnaire consisting of the case, and boxes to check for questions Q1-Q5. All cases in each data set are assessed in this way. A subset of cases, 10%, are double-marked to check for consistency.

4 Results

Recall and precision results for all queries and all data sets are given in Table 3. Due to the sample size, i.e. the amount of data recorded, we do not expect many statistically significant results. We begin by considering the hypotheses specified above.

Hypothesis 1 claims that precision and recall would be better for ontology-based matching than for keyword matching. The DS1 results confirm this claim.

Hypothesis 2 claims that precision and recall will be better than 90% for ontology-based matching. Precision is 100% for all queries on DS1, but recall is less than 90% in 2 out of 5 cases (the average recall is 85.6%). Consequently, only the part of the claim relating to precision is confirmed.

Hypothesis 3 compares recall and precision on new data sets. In 5 of the 6 queries posed to DS2 and DS3, recall is better for ontology-based matching than for keyword-based matching (mean recall rates 90.3% and 82.3% respectively). Precision is equal for the two techniques in 5/6 queries, with keyword-based improving on ontology-based matching in the remaining case. The claim for improved recall rates is confirmed, while the precision claim is rejected.

Hypothesis 4 is confirmed for precision in all cases, and in 8 of the 10 cases recall is greater than 80%. The claim is confirmed for precision rates.

Query No.	DS1		DS2		DS3	
	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)
Q1 ONT	100	100	100	100	100	100
Q1 KWD	100	100	100	100	100	100
Q2 ONT	70	100	89	100	88	88
Q2 KWD	40	100	44	100	75	100
Q3 ONT	100	100	75	100	90	100
Q3 KWD	100	100	75	100	100	100
Q4 ONT	94	100	83	100	88	100
Q4 KWD	-	-	-	-	-	-
Q5 ONT	64	100	90	100	48	92
Q5 KWD	-	-	-	-	-	-

Table 3. Recall and Precision Results

The variation factor is demonstrated if the comparative performance of matching techniques can be shown to vary between queries. Comparison of Q1 and Q2 for DS1 shows such a variation. Q3 turns out not to distinguish between matching techniques, that is, Q3 duplicates the variation effect with Q2 rather than duplicating the effect with Q1 (as was expected). The results are consistent across data sets.

5 Analysis

The hypotheses about precision rates are confirmed, but those on the absolute values of recall are not. The observed properties of high precision rates but variable recall rates were not foreseen. High precision comes from the use of unambiguous terms in keyword matching, or in the lexical terms associated with concepts. Variation in recall is traceable to two sources: the tokenising of the input and the use of unknown terms/concepts in the data. The unseen data used symbols such as `/ -` as word delimiters and these are not tokenised correctly by CB-IR (indeed, it may not be possible to account for this). The keyword approach did not tokenise the text so avoided the problem. Data set 3 contained many references to *Valves* which were judged to be a type of *Connector* hence relevant to Q5. However, the ontology does not contain this concept and none of these records were actually recalled by CB-IR. This explains the low recall rate for Q5 in DS3.

Statistical analysis of the data using a small sample test (Students t-test) did not find any significance in the mean recall rates either within a data set or for queries across data sets. For information, the mean recall rates for DS1, DS2 and DS3 are: 85.6, 87.4, 82.8, the variance of each of these values is comparable (but too high to yield significant results).

6 Conclusions

The empirical evaluation of ontology-based retrieval in CB-IR has broadly confirmed the hypotheses about relative and absolute performance of the system and about the adequacy and robustness of the ontology. A number of contributing hypotheses are not proven, but

the evidence tends to support rather than refute them. The hypotheses support the goal of demonstrating the technical utility of the system, which we consider to be achieved (with the qualifications stated above).

The evaluation has revealed an unexpected difference between recall and precision rates, and analysis of where errors arise has highlighted design issues where the system can be improved.

The Goal-Question-Metric approach proved to be a useful organising framework. However, the time and resources required for the evaluation are significant. As the ontology was one of the main subjects of the evaluation, it is appropriate to compare ontology acquisition and evaluation times. The ontology was constructed in several stages, and refined over a period of time, so it is difficult to estimate the development time accurately: 3-5 days is our best estimate. Evaluation has taken at least 5 days, and is therefore of same order of magnitude as ontology acquisition. The recording of human and system decisions was done manually and as a consequence was very time consuming. We conclude that accounting for evaluation by providing automated support for data collection and analysis is an important issue. In the case of information retrieval systems, it would be practical to built this capability into the system itself.

REFERENCES

[1] Cohen, P. *Empirical Methods for Artificial Intelligence*, MIT Press, 1995.

[2] Cohen, P. R., Chaudhri, V., Pease, A., and Schrag, R. Does Prior Knowledge Facilitate the Development of Knowledge-based Systems ? *Proceedings of The Sixteenth National Conference on Artificial Intelligence (AAAI-99)*
http://projects.teknowledge.com/HPKB/Publications.html

[3] Gomez-Perez, A. Evaluation of Taxonomic Knowledge in Ontologies and Knowledge Bases. *Proceedings of KAW'99*
http://sern.ucalgary.ca/KSI/KAW/KAW99/papers.html

[4] Menzies, T. Evaluation Issues with Critical Success Metrics. *Proceedings of KAW'98*
http://ksi.cpsc.ucalgary.ca/KAW/KAW98/KAW98Proc.html

[5] Menzies, T. hQkb- The High Quality Knowledge Base Initiative (Sisyphus V: Learning Design Assessment Knowledge). *Proceedings of KAW'99*
http://sern.ucalgary.ca/KSI/KAW/KAW99/papers.html

[6] Menzies, T. Web pages on evaluation.
http://www.cse.unsw.EDU.AU/ timm/pub/eval/
http://www.csee.wvu.edu/ timm/banff99/

[7] Nick, M., Althoff, K., and Tautz, C. Facilitating the Practical Evaluation of Knowledge-Based Systems and Organizational Memories Using the Goal-Question-Metric Technique. *Proceedings of KAW'99*
http://sern.ucalgary.ca/KSI/KAW/KAW99/papers.html

[8] Tallis, M., Kim, J., and Gil, Y. User Studies of Knowledge Acquisition Tools: Methodology and Lessons Learned. *Proceedings of KAW'99*
http://sern.ucalgary.ca/KSI/KAW/KAW99/papers.html