Extracting and Searching Knowledge for the Aerospace Industry

Vitaveska Lanfranchi[2], Ravish Bhagdev[2], Sam Chapman[2], Fabio Ciravegna[2] and Daniela Petrelli[1]

[1] Department of Information Studies, University of Sheffield,

[2] Department of Computer Science, University of Sheffield,
Regent Court, 211 Portobello Street, S1 4DP Sheffield, UK
{v.lanfranchi, r.bhagdev, s.chapman, f.ciravegna, d.petrelli }@shef.ac.uk

1. Introduction

A fundamental shift is occurring in many industries away from the selling of products (e.g. cars) to the provision of services (e.g. transport, car leasing). Essential to the long-term success of businesses in this emerging global environment is the creation of new Integrated Products And Services (IPAS). For example in Rolls-Royce plc (RR) the business model is changing from one of selling products (e.g. aircraft engines) and spare parts to one (driven by customer demand) of selling services (e.g. Total Care). This requires knowledge transfer between three very different worlds: new service design, new product design, and the operation of existing products and services in the field: the new product designer and the new service designer require significantly increased access to data on the behaviour of existing products in the field.

The IPAS project addresses these issues by integrating the three worlds: the separation of these worlds by geography, organisation, culture and time (decades), and their different information needs, make their integration very challenging. The objective of IPAS is to develop and exploit technologies such as meta-data, semantics, ontologies, text mining, search, social interactions, knowledge representation and semantic web services to enable the right information to be provided to the right person in the right form at the right time[1].

Knowledge is often stored in an unstructured format[2]. Textual documents cannot be queried in simple ways and therefore the contained knowledge can neither be used by automatic systems, nor be easily managed by humans. This means that knowledge is difficult to capture, share and reuse among employees, reducing the company's efficiency, effectiveness and competitiveness. For example, Service Representatives (SR) in RR create an Event Report (ER) every time a finding is recorded on a jet engine during service. Such information is unstructured (i.e. it is contained in an arbitrarily formatted Word file) but is very relevant to both designers and service representatives in order to gauge the problems experienced by the customers during service. As the information unstructured, the only way to access it is to use keyword matching.

Keyword matching systems are not very useful in this scenario because they only return documents that are likely to be relevant to a query. The kind of support users need is to receive data aggregated by content. For example, they need to produce statistics of problems and their causes, identify the components which are critical either because they often fail or because they cause disruption of service. Finding relevant documents is indirect way to access needed knowledge, as it then requires reading all the documents in order to extract the aggregated data. In this paper we describe how Information Extraction from text has been applied to ERs and how this extracted knowledge has been made available to users through an innovative search system for accessing the knowledge in a more direct way.

2. Information Needs in Aerospace Engineering

As mentioned previously, every time a Rolls-Royce jet engine is serviced in any airport around the world a report (Event Report, ER) is written by a SR and submitted to the control centre. While currently this information is remotely archived in a database by SRs, until recently ERs were sent as email attachments (MS Word files) to the control center. ERs are usually very short documents (about one page) that contain key information on the event (generally in tabular forms) such as engine type and number, airline operator, location, event description and actions taken, etc., plus a short natural language text describing the event.

The history of each single engine and its component parts is captured in a series of ERs. When searching for information inside legacy ERs, no special search feature is provided other than the standard MS Windows® search mechanism so several search steps as well as manual work is necessary for filtering the results. For example service engineers in the customer service unit can be interested interested in monitoring the fleet and minimising the impact of maintenance on flight schedules; the history of the engines is therefore assessed to determine which situations need attention. If an engineer is interested in knowing which past events have caused: 1) a flight delay or cancellation, and 2) required the installation of a new Fuel Metering Unit (FMU); and 3) a fuel leak was discovered, several steps need to be carried out in order to get a satisfactory answer.

Service engineers are not the only user group interested in ERs. Designers involved in the planning of new engines are interested for example in discovering how the component or the part they are responsible for deteriorate and wear during use. For instance, a designer of a FMU might often need to ask questions such as "What events resulted in replacement of FMU on an engine type PQ123?".

---

[1] www.3worlds.org

[2] According to Prabhakar Raghavan (Yahoo inc.), an average 80-85% of a company's knowledge is contained in unstructured form, i.e. expressed in unstructured form, e.g. in natural language

In the example, service engineer's and designer's requests both concern FMU, and the same data (past ERs) are then analysed with different perspectives and intentions of use. Both types of users perform recall-oriented search: where it is essential that all instances are retrieved. As mentioned above, currently both service engineers and designers spend considerable time searching and reading ERs. However, despite their (often extended) effort, they may end up with just a handful of documents as current technology does not guarantee the retrieval of all the relevant ones. At the same time precision plays an important role: presenting only relevant search results reduces the users time in completing their task. The most urgent need for RR (Rolls-Royce) users is then of a search engine mechanism supporting precise, focused and effective recall-oriented retrieval, allowing complex queries and accommodating different perspectives.

To better understand the users' activity and [3] and derive user requirements [4] a set of user studies (encompassing a questionnaire, interviews and observations) has been carried out in RR. The main requirements were:

- The system should help the user in quickly focusing on goal of the search;
- Users have complex information needs and the system should allow them to express complex queries in a simple way;
- The best searching strategy depends on the task: the user should be able to quickly change their research strategy or focus;
- Different users may use different terminology to refer to the same object; the system should accommodate this individual perspective;
- RR users usually plot the results in graphs using external tools: the system should automatically perform this step and graphs should be generated on demand; more specifically:
    - It should be possible to manipulate the charts, e.g. changing the dimensions and the grouping of the items, in order to reflect alternative views on the retrieved data.
    - Each chart component should be the interactive means to further inspect a subset of the retrieved data.

In the next section we will first of all describe the generic methodology devised to answer the requirements and then the specific system, X-Search, implemented for the use case.

3. Hybrid Search for Event Reports

As mentioned, ERs are generally short and the language very specialised, two conditions critical for traditional keyword-based retrieval. Past research on retrieving very short text (e.g. image captions [2]) has shown traditional techniques fall short to be effective; moreover technical documents that use very limited vocabularies have proved to be challenging for keyword-based retrieval (e.g. in car manufacturing [1]).

A further complication technical text presents for traditional IR techniques is the context in which relevant keywords occur. In the example "find all cases in which a blade was changed due to corrosion, the fact that the corrosion was found on the blade is fundamental to answering the question. A search engine that retrieves all documents where the terms corrosion and blade co-occur would retrieve too many irrelevant ERs to be of any practical use. Indeed aerospace engineers are not interested in the relevance of the ER per se (i.e. number of documents), but on the knowledge that they provide, that is to say on the cases where corrosion was on the blade.

Ontology-based indexing and Semantic Web (SW) technologies can be used to associate formal metadata to text, making the document content (as opposed to its keywords) available for automatic processing [2]. An ontology is used both for annotating the documents and for searching by concepts; it allows linking of synonyms to the same concept (name, acronym and number all referring to the same part) or relate concepts through logical statements (corrosion on the blade). Search based on metadata does not suffer from any of the problems mentioned above for keyword-based searching, as it is uninfluenced by the length of the text or on the distribution of words in it.

Despite its power, a SW approach has limitations as it constrains the search to the information captured accordingly to the ontology. In our experience with real-world applications, the ontology initially developed often does not cover the full user needs as the use of information takes forms unexpected at ontology design time and modifications to the ontology are complex and expensive hence seldom performed. Moreover, the generation of metadata can be expensive if done manually or error prone if done automatically. Finally, queries must be formulated in a logical language, which is generally considered very difficult by normal users.

Conversely to semantic techniques, keyword-based information retrieval has the advantage of being flexible - any term can be searched independently from previous processing - and straightforward to use, just type terms. In this paper, we claim that a hybrid approach that unites keyword based and ontology-based search is able to combine the advantages of both techniques, providing effective, flexible and focused search that classic methods alone cannot achieve.

Hybrid Search (HS) combines the flexibility of keyword-based retrieval with the ontology and its reasoning capabilities, making a synergistic use of both strength, supporting the user in focusing on relevant issues with faster and more accurate results. In HS, users can combine within the same query: (i) ontology-based search; (ii) keyword-based search and (iii) keyword-in-context based search. Keyword-in-context searches the keywords only in the text previously annotated with a concept in the ontology; for example searching for "fuel" in the context of the removed parts listed in an ER. HS is enabled by three offline steps: (i) indexing documents using keywords, (ii) defining a domain ontology and (iii) extracting information from documents using an ontology.

4. X-Search

The HS General approach has been implemented in the X-Search system. This system is currently under final test at Rolls Royce in Derby (UK) to access ERs by both service engineers and designers.

In order to extract information from the ERs, an existing ontology describing (among other things) engines parts and event related metadata (e.g. location, author) was selected; the ontology was built independently by the University of Aberdeen as part of the IPAS project.

Information extraction was then performed on the tabular part of ERs only. An extractor was developed that uses Support Vector Machine approach to learn over the documents. This was developed as a plug-in for T-Rex[3]. Once the information is extracted, it is stored in the form of RDF triples in a triple store[3]. Indexing documents using keywords was performed with a standard indexing system (Nutch[4]).

At retrieval time, X-Search performs the following steps:

- The query is parsed and the types of searches are identified (keywords, keywords-in-context and ontology-based) and separated;
- Keywords are sent to the traditional information retrieval system; this will return the identifiers (URIs) of all the documents that contain those keywords;
- Queries about concepts (and their relations) are matched with the facts in the knowledge base using a query language like SPARQL[5];
- Queries of keywords-in-context are sent to the knowledge base, returning conceptual instances containing the given keywords (again using SPARQL);
- Finally, the results of the different queries are merged.

The query interface (see Figure 1) enables users to perform both ontology-based and keyword-based queries, as well as a combination of the two.

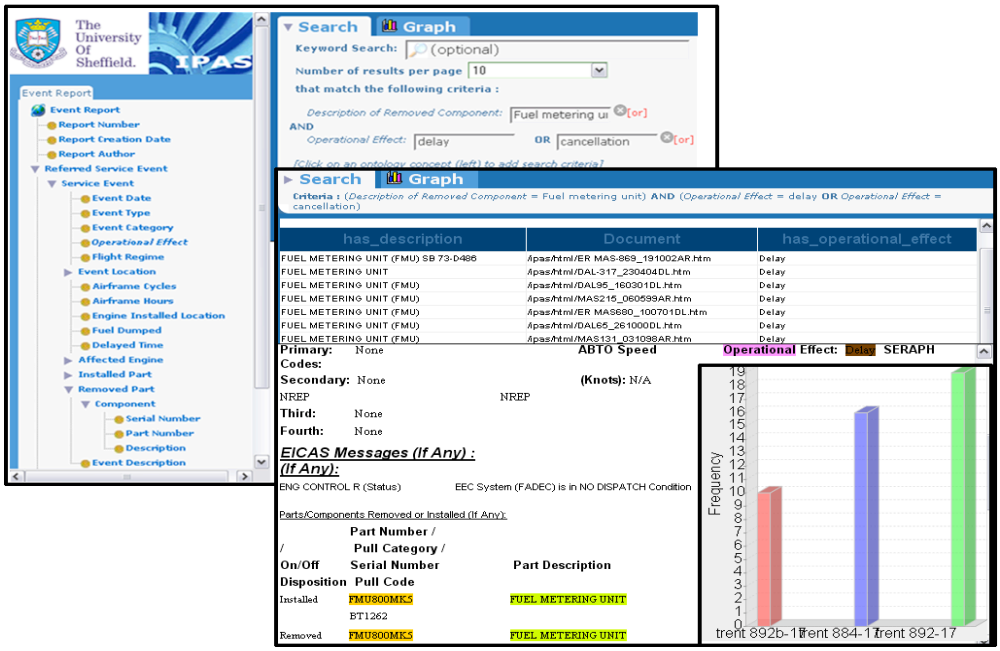

Figure 1 – X-Search querying interface and visualization of results

It is possible to query the archive by:

- Strict ontology-based queries: a precise description must be provided in the query. For example it is possible to query for events happening in the UK and receive events which happened in Manchester, London, etc.
- Ontology-based keyword matching: it is possible to apply keyword matching on the descriptions identified as belonging to a specific type. For example it is possible to retrieve all the documents where the removed part contains the word "fuel". This is useful because it enables partial matching on the description in case the user wants to input a less precise query but still make use of the structured knowledge.
- Plain keyword matching, where the keyword can appear anywhere in the document.

[3] Sesame (http://www.openrdf.org/)
4 Lucene (http://lucene.apache.org/nutch/)
[5] SPARQL (http://www.w3.org/TR/rdf-sparql-query/)

When a query is performed, the result set contains the ERs where the concepts and the keywords in the query co-occurr. The set is displayed as a list on the mid-right panel of the interface; each item in the list has the name of the document and the values of the fields used for ontology-based search. Individual ERs are shown on the bottom right when requested (clicking on a list item). Multiple documents can be opened simultaneously, each one displayed in a different tab.
The original layouts of the documents are maintained while they are converted to HTML format (see Figure 1 for example). Annotations are made evident through colour highlighting (as in [5,6] and are the means to advanced features or services [6,7]: for example clicking on a concept generates a query expansion with the selected term.
One of the identified user requirements is to provide the automatic quantitative analysis of the retrieved set and create graphs and charts to summarise it.
X-Search provides the user with the possibility of choosing the style of the graph and the variables to plot. The graph in Figure 1 plots the results of the previous query by engine type. Each graphic item (each bar in the example) is active and can be clicked to focus on the sub-set of documents that contains that specific occurrence.

5. Evaluation

In order to prove the effectiveness and utility of the X-Search system, a set of experiments where run. First of all the quality of the metadata generated by the Information Extraction was evaluated testing the effectiveness of the T-Rex [3] plugin on the annotated corpus of 400 documents. The set of documents was divided into training and test sets (using 50% approx. split) and the learning curve studied. As expected, the system performance improved as the training set size increased. For example, for the concept ”Part Installed Description”, when 40 documents were used for training, Precision (P) was 76.00% and Recall (R) was 100.00% while with 240 documents P increased to 90.22% (R remained 100.00%).

The combined evaluation results on all field's obtained in a two-cross folder test using 50.00% of the corpus, were Precision=95.12%, Recall=97.00%, F-Measure=95.84%. This shows that the Information Extraction system is very good at generalising over the differences in table formatting, despite their irregularity. The quality of the extracted metadata was proven to be very high and therefore we could proceed to test the effectiveness of the hybrid search without risking it to be adversely affected by metadata quality.

A comparative approach was adopted to evaluate the HS with respect to keyword and ontology-based search. The goal of the evaluation was not to demonstrate that the HS is more powerful than the other two, but instead to understand if and when the combination of the two provides an advantage in focusing the search and reducing the burden on the user side. Indeed X-Search (as discussed in Section 4) offers the user all three modalities, to select the most effective for the task at hand.

A set of 21 topics was generated on the basis of observed tasks, sequences of user queries recorded in the RR corporate DB or as elaboration of direct input from RR users (i.e. examples of their recent searches). Some topics, like "How many events were caused during maintenance in 2003?", can be answered using ontology-search alone, others, like "What events were caused during maintenance in 2003 due to control units?" by combining annotations and keyword. Finally one topic, i.e. `Find all the events associated with damage to acoustic liners following bird strike", can only be answered using keyword-based search.

The effectiveness of keyword-based search, ontology-based search and HS was tested in all the 21 cases. For the topic that required keyword search to be answered the goal was to maintain the same accuracy in the HS. The goal of the other two sets was to test how the three methods compared. In order to run the evaluation, topics were transformed into queries by selecting the corresponding concepts or composing the adequate query terms. A pool of retrieved documents was built collecting all the results of the runs and manually assessed for each topic to determine the relevance of each document in the pool (binary relevance was used, i.e. relevant or non-relevant). The set of relevant documents (**POS** in the following) was then used to measure Precision and Recall at 20 and 50 (using the first 20 and 50 hits returned for each query respectively). Precision was calculated by computing the number of correct hits divided by the results returned:

$$Precision = \frac{COR}{min(ACT, maxNo)}$$

where ***maxNo*** is either 20 or 50, COR is the number of correct hits returned by the system and ACT the number of returned documents.

To compute Recall: $Recall = \frac{COR}{EXP}$ where ***EXP=min(POS, maxNo)***.
The reason for the definition of EXP is to avoid penalising the system for not returning more than ***maxNo*** documents when checked on ***maxNo*** documents (e.g. if there are 100 relevant documents and just the first 20 hits were checked, then 20 was considered the number of relevant documents). The HS's effectiveness was computed in two ways (Figure 2): HS Strict and HS General. HS Strict was applied when there was only the possibility of performing a true hybrid search, i.e. when the query had both an ontological and a keyword part. HS General is the application of HS Strict plus the application of the best of either keyword or ontology-based search when strict was not applicable. HS General is the strategy that we have implemented in the X-Search system and that we consider the true form of HS.

Ontology-based search has very high precision, but the lowest recall. This is because the ontology did not model 6 of the topics. Keyword search has lowest precision and fairly good recall. HS Strict has the highest precision, but low recall, due to the fact that 5 topics did not require HS Strict. HS General reports very high precision (1% worst than ontology-based,

+51% with respect to keywords), and the highest recall (+46% with respect to keywords and +109% with respect to ontology-based search). F-Measure is +49% with respect to keywords and +55% with respect to ontology-based. HS General reports -2% in Precision and +81% in Recall with respect to HS Strict (F-Measure is +40%).

In conclusion, HS General experimentally outperforms all the other methods.

Query	POS	Keyword 20			Ontology 20			Hyb 20 Strict			Hybrid 20 General		
		COR	ACT	EXP	COR	ACT	EXP	COR	ACT	EXP	COR	ACT	EXP
Q1	84	16	20	20	20	20	20	0	0	20	20	20	20
Q2	22	16	20	20	0	0	20	7	7	20	16	20	20
Q3	25	1	20	20	11	20	20	0	0	20	11	20	20
Q4	63	19	20	20	19	20	20	0	0	20	19	20	20
Q5	27	9	20	20	12	20	20	0	0	20	12	20	20
Q6	5	4	8	5	0	0	5	3	7	5	4	8	5
Q7	7	6	6	7	0	0	7	4	4	7	6	6	7
Q8	1	1	1	1	0	0	1	1	1	1	1	1	1
Q9	5	3	3	5	0	0	5	5	5	5	5	5	5
Q10	83	12	20	20	0	0	20	20	20	20	20	20	20
Q11	2	1	1	2	0	0	2	1	1	2	1	1	2
Q12	3	3	3	3	0	0	3	3	3	3	3	3	3
Q13	7	6	6	7	0	0	7	6	6	7	6	6	7
Q14	145	19	20	20	19	20	20	20	20	20	20	20	20
Q15	40	8	20	20	0	0	20	20	20	20	20	20	20
Q16	11	1	16	11	11	11	11	0	0	11	11	11	11
Q17	13	3	20	13	0	0	13	4	4	13	4	4	13
Q18	7	1	4	7	0	0	7	4	20	7	4	20	7
Q19	25	10	17	20	0	0	20	11	11	20	11	11	20
Q20	53	3	20	20	20	20	20	0	0	20	20	20	20
Q21	37	18	20	20	0	0	20	20	20	20	20	20	20
TOTAL	665	160	285	281	112	131	281	129	149	281	234	276	281
		PREC	REC	F-MEAS	PREC	REC	F-MEAS	PREC	REC	F-MEAS	PREC	REC	F-MEAS
		0.56	0.57	0.57	0.85	0.40	0.54	0.87	0.46	0.60	0.85	0.83	0.84

Figure 2 - Evaluation results

6. Conclusions

In this paper a hybrid search approach has been proposed as a methodology for the analysis of ERs from RR corporate archives and implemented in the X-Search system. This approach extends the structure and reasoning of semantic search paradigm combining it with the flexibility and expressivity of keyword-based retrieval; among the advantages:

- using an ontology it is possible to overcome the problems of synonymity and abbreviations, as the ontology uniquely identifies objects;
- the metadata can be used to model the context in which the information is captured via ontology-based logical statements;
- different user perspectives are taken into account;
- the search results are more focused and precise than traditional methods; gain in terms of precision and recall with respect to keyword based searching is in the order of 40/50%, while the gain in terms of recall with respect to ontology-based search is in the order of 100% with an equivalent precision.
- the results can be automatically plotted in graphs.

X-Search was designed and developed taking into account RR business and user needs, facilitating searches through corporate archives. As a result, the search activity can be performed in a faster and more effective manner by both designer and service community. Moreover the hybrid search approach avoids the costly need of modifying a highly structured ontology and still allows users to not be constrained by the scope of the structured knowledge. A formal user evaluation is currently being undertaken in RR in Derby (UK). It will assess the usefulness and ease of use of the X-Search system and its interaction approach. This phase will be followed by a controlled deployment of the X-Search system to final users.

Acknowledgements

The authors would like to thank Colin Cadas of Rolls Royce Plc for his invaluable help and assistance in both the project work and also this publication. IPAS is part funded by the Department of Trade and Industry under the Technology Program and by Rolls-Royce plc. DTI Reference TP/2/IC/6/I/10292

References

[1] Fabio Ciravegna, Understanding messages in a diagnostic domain, , Inf. Process. Management, 31, (5), 1995, 0306-4573, 687—701, Pergamon Press, Inc., Tarrytown, NY, USA

[2] Tim Berners-Lee, James Hendler and Ora Lassila, The Semantic Web, Scientific American, 2001.

[3Jose' Iria and Fabio Ciravegna, A Methodology and Tool for Representing Language Resources for Information Extraction, 5th International Conference on Language Resources and Evaluation (LREC 2007), Genoa, May, 2006.

[4] Daniela Petrelli, Vitaveska Lanfranchi, Phil Moore, Fabio Ciravegna and Colin Cadas, Oh my, where is the end of the context? Dealing with information in a highly complex environment, IIiX: Proceedings of the 1st international conference on Interaction in context, 2006, 1-59593-482-0, 37—41, Copenhagen, Denmark, ACM Press, New York, NY, USA

[5] Fabio Ciravegna, Alexiei Dingli, Daniela Petrelli and Yorick Wilks, User-System Cooperation in Document Annotation Based on Information Extraction, London, UK, EKAW '02: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, 122--137,Springer-Verlag, 2002

[6] Martin Dzbor, John Domingue and Enrico Motta, Magpie - Towards a Semantic Web Browser, Second International Semantic Web Conference, Sanibel Island, FL, USA, October, 2003, 2003, 690-705

[7] Vitaveska Lanfranchi, Fabio Ciravegna and Daniela Petrelli: Semantic Web-Based Document: Editing and Browsing in AktiveDoc, Proceedings of the 2nd European Semantic Web Conference (ESWC 2005), 623-632, Heraklion, May 2005