Doris: Managing Document-based Knowledge in Large Organisations via Semantic Web Technologies

Ravish Bhagdev, Jonathan Butters, Ajay Chakravarthy, Sam Chapman,
Aba-Sah Dadzie, Mark A. Greenwood, José Iria and Fabio Ciravegna.

Department of Computer Science, University of Sheffield,
Regent Court, 211 Portobello Street, S1 4DP Sheffield, United Kingdom
{N.Surname}@sheffield.ac.uk

Abstract. The acquisition, sharing and reuse of knowledge is a prime challenge in large organisations. Doris is a framework for defining Knowledge Management applications based on Semantic Web technologies that enables flexible means of capturing knowledge and of searching and exploring the knowledge and the documents where it is contained. Applications of Doris are employed in the aerospace domain at Rolls-Royce plc and in the domain of exploring and searching archives about London of the 18th century.

1 Introduction

The objective of this challenge is to build an environment to effectively acquire and share knowledge in large distributed environments. Knowledge is acquired from documents (either from legacy documents or while new documents are created) in the form of ontology-based annotations. Then both knowledge and documents are made available for searching and for qualitative and quantitative analysis. Requirements for such an environment are:

- The tool must support the user's everyday work and tasks with very limited intrusiveness. For example it must not ask to perform detailed manual annotation of documents, which would distract the users from their tasks.
- It must enable to accommodate different user requirements, tasks and profiles, accommodating, for example, different search and acquisition strategies.
- The information or knowledge provided to users must be very high quality; this is important as some of the technologies used (e.g. Information Extraction, IE, from texts) can perform poorly in some cases.
- It must be portable to different domains and tasks with limited effort and in a limited time.
- It must be able to work over a medium/large scale, coping with dozens (or hundreds) of thousands of documents and thousands (or millions) of triples; documents are expected to be multimedia objects containing text and/or images.
- It must be able to integrate knowledge, information and documents from distributed heterogeneous sources.

- It must be able to cope with a dynamic world where new information and documents are added constantly and must readapt to such changes.

2 Doris: effective knowledge acquisition, sharing and reuse

Doris is a framework for defining Knowledge Management (KM) applications based on Semantic Web (SW) technologies. SW technologies enable flexible means of (i) capturing knowledge from both users and documents, and (ii) searching and exploring knowledge and documents containing this knowledge. Its main features:

- It enables knowledge capture at document creation time with minimised intrusiveness for the user; traditional technologies like HTML form filling are employed to create documents, but SW technologies work in the background both for producing the most appropriate acquisition strategy and for capturing knowledge from user input; knowledge is captured both in text and images (and across both). Knowledge capture is also facilitated by the use of machine learning and IE from documents.
- It enables knowledge capture from large amounts of legacy documents using ontology-based IE from texts; this is quite important as large organisations generally have hundreds of thousands of historical documents containing very valuable knowledge.
- It enables user-centric searching of documents and knowledge using hybrid search, a strategy that accommodates ontology-based and keyword-based searching in a flexible way. Searching and sharing is adapted to the user's needs.

The architecture is fully based on open Web and SW standards. The central means of representation and communication of knowledge are (a set of) ontologies that define both the domain and the user tasks.
In the following sections we will introduce Doris' main features.

2.1 Capturing Knowledge at Creation Time

One of the main issues in KM is capturing knowledge when it is created. In traditional KM models, knowledge is stored in unstructured documents (e.g., Word files) or in databases. The problem with collections of unstructured documents is that knowledge sharing and reuse can only be performed via keyword-based searching and that knowledge in documents may be left implicit, and therefore unavailable. Database schemas and associated form filling methods are instead very inflexible and generally do not support users with different requirements.
In Doris, knowledge can be captured at creation time, by:

- Generating exhaustive and explicit information without forcing users through a pre-defined set of steps. Capturing strategies can be flexibly and declaratively designed for the specific user's needs, to the point that every user can have a different acquisition strategy.
- Capturing and classifying information and knowledge via an ontology, so to make it available for sharing and reuse.

For this purpose, we have designed a component that enables easy definition of strategies for user-centred knowledge capture via declarative means. AktiveForm is designed to provide controlled domain/task-specific document creation and ontology-based annotation of the contained knowledge. AktiveForm uses two ontologies: a user ontology, to define the user tasks and processes, and a domain ontology, to tag the knowledge inserted by the users.

The user ontology allows for flexibility in supporting the users, as it provides a way to describe the kind of support the individual user may need: for example, as different users can have different knowledge (e.g. an expert can have more sophisticated knowledge than a man on the street), the knowledge capturing process can be tuned to the capabilities of each. The resulting form will be custom tailored, offering different types of fields or different features. This is achieved by declaring in the ontology also the required information, relevant restrictions and how to acquire it based on the user profiles. Relations among fields such as precedence, layout and services associated are also established. A reasoner then uses this information to plan the actual form, the resulting document and the acquisition process. The interface displays the form using AJAX.

Data inserted by the user is automatically tagged via the domain ontology, making it available for semantic searches (see section 2.3.1).

AktiveForm enables the input of data via many interface strategies, like drop down menus for enumerated fields, free text fields and free text forms. Value checking is performed using the ontology and the reasoner. For single text fields (e.g., to input the description of a jet engine component), support by a terminology recogniser is provided to identify the correct value: this is particularly important for fields with hundreds of possible values (e.g. there are 300,000 parts in a jet engine). In this case the user can input a rough description and the terminology recogniser (constrained by the expected value) proposes the correct term (or a set of possible ones). Terminology recognition is based on String Distance Metrics [1].

Free text fields containing unstructured descriptions are allowed, and their content can be semantically annotated either by the user (using an annotation tool such as AktiveMedia [2]) or automatically by an IE module (see section 2.2). Images can be added and annotated using the same strategy.

When a document has been created by filling in all parts of the form, a PDF file is generated containing the whole generated document. Different documents can be generated for different users according to the intended usage, by composing the extracted information and knowledge. Again the ontology and the reasoner explicitly control the generation of the document and its layout.

2.2 Acquiring Knowledge from Legacy Documents

A large amount of unstructured legacy documents (e.g. Word documents, PowerPoint presentations, etc.) is available in large organisations. We have mainly focused on textual documents (e.g. Word or HTML files), containing images. The acquisition architecture is illustrated in Fig. 1; documents are either discovered by a

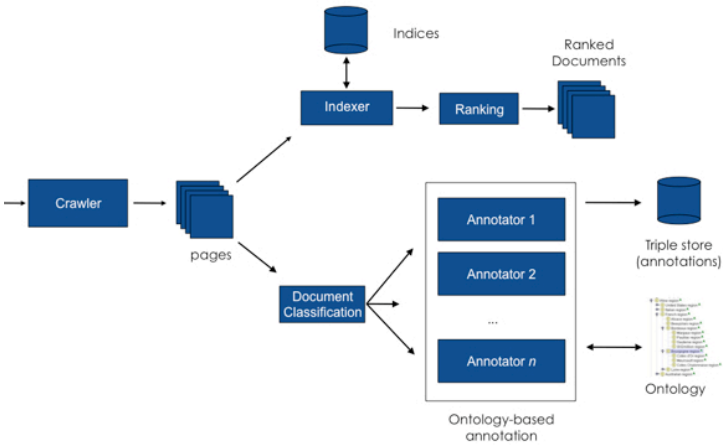

Fig. 1: Document indexing and annotation: traditional keyword indexing and document ranking (top of figure) is done in parallel with ontology-based annotation (bottom).

crawler or submitted individually to the system and then indexed using a traditional keyword based indexer (plug-ins are available for Lucene and Nutch), which will generate inverted indexes.

In order to extract knowledge and make it available to other components within Doris, documents are classified and the relevant ontologies for annotation identified. This can be done manually by the user (e.g. by assigning directories or collections of documents to a specific task/ontology) or automatically by a trained classifier. Then the document is annotated using one of the available plug-ins, i.e., T-Rex[3] or Saxon, a rule-based extraction system (or any combination of the two). Doris also allows for manual annotation (or correction of automatic annotation) via the AktiveMedia module [2]. Extracted information (ontology-based annotations) is stored in the form of RDF triples according to OWL or RDF ontologies in a triple store (currently Sesame or 3Store).

Doris uses an ontology-based data representation model that greatly enhances the extensibility of its knowledge acquisition modules. All the modules use the same underlying representation for the input documents by referring to the same ontology that models how the documents should be represented for the purposes of knowledge acquisition. The ontology specifies which entities, and which relations between those entities, should be instantiated in the representation when documents are processed by NLP tools such as tokenizers, part-of-speech taggers and parsers. Modules are plug-ins managed within a framework [4] that automatically decides which modules to run based on the data representation model chosen. This has allowed the system to be extended and ported to a number of domains via a definition of the appropriate ontology that models the data to be obtained from the documents in that domain.

2.3 Knowledge Sharing and Reuse

As the goal of knowledge management is to provide the right information at the right time, searching can be considered a main way of sharing.

2.3.1 Searching

Searching is performed using X-Search, a component within Doris that implements a Hybrid Search (HS) strategy, mixing keyword-based searches (KS) and ontology-based searches (OS) [5]. HS combines the flexibility of full text keyword-based retrieval with the ability to query and reason on document metadata. In HS, users can combine, within the same query: (i) OS via unique identifiers (e.g. *URIs* or unique identifiers); (ii) KS and (iii) keyword-in-context. Keyword-in-context searches the keywords only in the portion of the document annotated with a specific concept in the ontology; for example in an aerospace domain, it enables searching for the string "*fuel*" but only in the context of all the text portions annotated with the concept *affected-engine-part*.

At retrieval time, HS requires the following steps:

- the query is parsed and the three types of searches 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 containing those keywords; performing two types of matches: strict matches, where all keywords must be present in the returned documents, or less strict matches where some of the keywords can be missing from the documents. X-Search uses Nutch for indexing documents, because of the high quality keyword mechanism of Nutch and its ability to exploit ranking strategies used by search engines.
  - queries about concepts (and their relations) are matched with the facts in the knowledge base using a query language like SPARQL[1]; results can be returned that strictly match the results; in a more sophisticated approach it is possible to perform near matches, for example by automatically relaxing constraints. Concerning support for triple stores, X-Search provides plug-ins for Sesame and 3Store; query languages supported are SPARQL and Sesame's SeRQL.
  - queries of keywords-in-context are sent to the knowledge base, returning conceptual instances containing the given keywords (again using SPARQL); again near matches can be performed;
- Finally, the results of the different queries are merged, ranked and displayed.

A user query is created via a web form interface that enables easy graphical composition of ontology and keyword-based conditions (left part of Fig. 2). Keywords can be inserted into a default form field in a way similar to that required by search engines; Boolean operators AND and OR can be used in formulating queries. Conditions on the metadata can be added to the query by clicking on an ontology concept.

2.3.1 Quantitative Analysis of Information (graphs)

Quantitative analysis of information enables users to explore the knowledge space and then drill down to documents of provenance to check the details. X-Search supports quantitative analysis by automatically generating graphs and charts. The user can choose the style (pie or bar chart) and the variables to plot. Each graphic item is

[1] www.w3.org/TR/rdf-sparql-query/

active and can be clicked to focus on the sub-set of documents that contains that specific occurrence. Both graphs and documents can then be shared with other users.

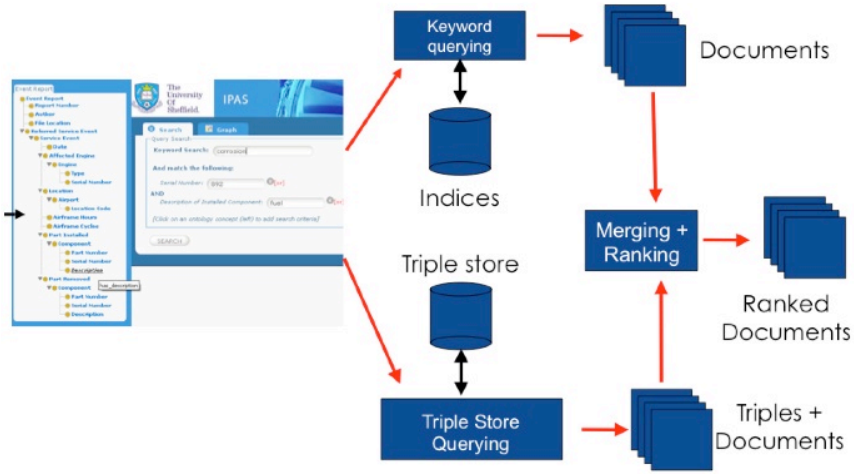

Fig. 2: Combining keywords and ontology-based searching in hybrid search.

3 Testing and User Evaluation

AKTiveForm has been tested informally in-house and is currently undergoing further development prior to more comprehensive pilot tests and user evaluation. This will address interface limitations such as moving between sections of a form during fill-in.

X-Search has undergone in-house stress testing, in addition to evaluation by real users in their working environments, where it emerged that some users first performed keyword-based searches and in a second iteration added conditions on the ontology; others composed conditions on the ontology and keywords in a single step; yet others used the ontology as a first approach and added keywords later to refine the search. HS overcomes some of the limitations of the methodologies used for knowledge capture, and in particular IE from legacy documents. It is a well known issue with IE that in some tasks (e.g. complex event capturing) there is an upper limit in its accuracy (defined in the MUC conferences [7] as the 60/70 Precision/Recall upper limit). In Doris, IE is designed so as to perform at a very high accuracy (above 90% in terms of precision/recall) to enable high quality ontology search. For the pieces of information where ontology-based annotations are not available this accuracy is unreachable. Searching however remains effective as the keywords used are constrained by the presence of other ontology-based constraints (confirmed during evaluation as users never complained about the limitation). Also, keyword search can be limited to specific sections of a document, further constraining search.

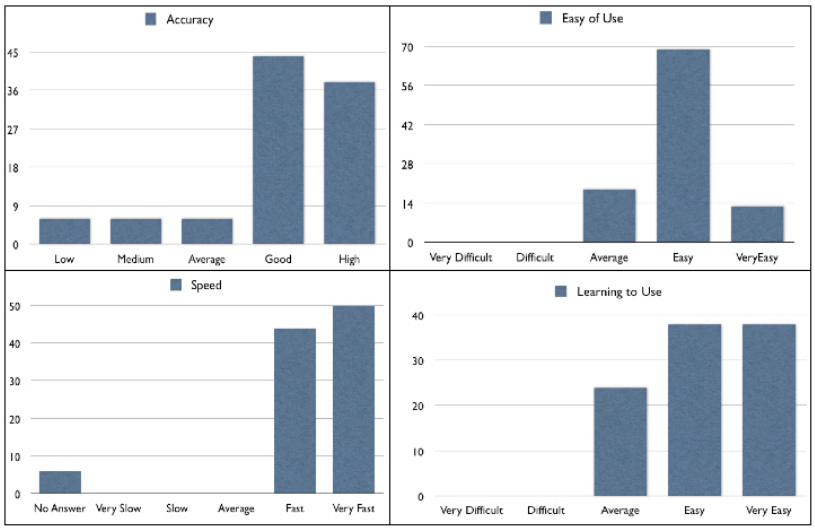

Fig. 3: Results of evaluation of X-Search by 32 users (values are in %).

4 Conclusions

In this paper we have presented Doris, a framework for the definition of document based Knowledge Management. It enables semantic knowledge capture both while documents are created and from legacy documents. It also enables flexible search across documents and knowledge. The interesting aspects of Doris are:

- All the applications work in a browser on client-server architecture; either locally or on a remote server. Therefore they can be used even in mobile situations and synchronised at a later time.
- Ontology-based annotations are generated as a side effect of user activity (via form filling, as fields in a form are associated to concepts in ontologies) or as a direct effect of explicit user-centred annotation. Annotations are also generated automatically for legacy documents by the IE modules.

The interesting aspects of knowledge capture at creation time are:

- The generation of applications is completely declarative and done via ontologies and reasoning; changes do not require any programming, just the modification of a simple ontology.
- The adoption of applications in a work environment does not require any workflow change for the users. The user workflow can be simulated by the form filling activity and appropriate non intrusive strategies can be defined. Again, defining a new class of users with different acquisition requirements does not require any programming, just changes to the ontology.
- At the end of the process, as well as structured knowledge a PDF file is generated which can be printed or sent out by email.

The interesting aspects of the knowledge acquisition from legacy documents are:

- Traditional keyword-based document indexing is done in parallel with ontology-based IE from the documents; this approach is also found in other systems such as KIM [6]; however, in Doris this is used at search time via HS.
- An ontology-based data representation model enables extensibility of the knowledge acquisition capabilities of Doris.

Concerning searching, Doris supports multiple approaches to searching and sharing through the HS framework, combining keyword only, ontology only or different combinations of both.

In order to test the effectiveness of Doris, two real world applications were developed:

- Company's Knowledge Management: a knowledge capturing application was developed for Rolls-Royce plc. Two corpora containing dozens of thousands of documents were used for extraction and the resulting system was released in an alpha test to 32 Rolls-Royce engineers (see Fig. 3 for user evaluation results). The system is currently under beta test by about a hundred users. In this application, we have so far analysed only legacy documents. The capturing of information at creation time is expected to enter alpha test in early autumn 2007 after the effectiveness on legacy data is fully assessed.
- Historical information search in distributed archives for London of the 18th century: this is a more limited application, which enables historians to navigate data about 18th Century London by simultaneously searching data in heterogeneous archives.

Acknowledgments. The work was jointly supported by IPAS, which is funded by the UK DTI (Ref. TP/2/IC/6/I/10292) and Rolls-Royce plc, and X-Media, (www.x-media-project.org), EC grant number IST-FP6-026978. The application to the history domain was sponsored by the AHRC. Thanks to Matthew Rowe, Lei Xia, Daniela Petrelli, Vitaveska Lanfranchi and Ziqi Zhang for helping in the preparation of this challenge.

References

1. Chapman, S. SimMetrics: a similarity library of metric algorithms for integration and comparison, 2004, http://www.dcs.shef.ac.uk/~sam/simmetrics.html
2. Chakravarthy, A., Lanfranchi, V., Ciravegna, F.: Cross-media Document Annotation and Enrichment, in Proc. of the 1st Semantic Authoring and Annotation Workshop, ISWC2006, Athens, GA, USA, 2006
3. Iria, J., Ireson, N., Ciravegna, F.: An Experimental Study on Boundary Classification Algorithms for Information Extraction using SVM. In Proc. of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006), at EACL 2006, April 2006.
4. J. Iria and F. Ciravegna. A Methodology and Tool for Representing Language Resources for Information Extraction. In Proc. of LREC 2006, Genoa, 24-25-26 May, 2006.
5. V. Lanfranchi, R. Bhagdev, S. Chapman, F. Ciravegna, D. Petrelli, Extracting and Searching Knowledge for the Aerospace Industry, Proceedings of 1st European Semantic Technology Conference, May 2007, Vienna, Austria.
6. Popov B., Kiryakov A., Ognyanoff D. , Manov M., Kirilov A. KIM - A Semantic Platform For Information Extraction and Retrieval, Journal of Natural Language Engineering, Vol. 10, Issue 3-4, Sep 2004, pp. 375-392, Cambridge University Press.
7. Message Understanding Conference (MUC)7, Linguistic Data Consortium, PA USA, 2001.