Managing Collaboration Projects using Semantic Email Search

M-Dyaa Albakour
University of Essex
and
Active Web Solutions
malbak@essex.ac.uk

Rob Blackwell
Active Web Solutions
rob.blackwell@aws.net

Udo Kruschwitz
University of Essex
udo@essex.ac.uk

Simon Lucas
University of Essex
sml@essex.ac.uk

Categories and Subject Descriptors

H.4 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

UIMA, email processing

1. INTRODUCTION

Natural Language Processing (NLP) and Semantic Web technologies have matured significantly in recent years. At the same time it seems surprising that we do not already have more applications that make extensive use of these methods. There are a number of reasons, e.g. proper natural language understanding is still an area of active research [4], whereas practical Semantic Web applications typically rely on significant knowledge construction efforts but "the Semantic Web lacks sufficient user involvement almost everywhere" [7].

The approach that we have taken is to demonstrate how NLP and Semantic Web techniques can be usefully combined to provide a practical real-life application with minimal manual customization. We are addressing a specific issue of project management often found in large cross-enterprise collaboration projects that involve a variety of partners. We observed that (not surprisingly) most information in such projects is exchanged via email. That makes it difficult to keep track of the state of the project, what has been agreed on and how different tasks within the same project relate to each other. We have also observed that in such contexts the project partners are reluctant to use existing extranet systems as it can be too much of an overhead.

The key objective of *Infoplaza* is to extract a knowledge base from the project mailboxes by semantically analysing emails using NLP and Semantic Web technologies. This knowledge base facilitates the task of automatically indexing and classifying email messages. The knowledge base also provides a way of semantically searching the accumulated knowledge, providing a snapshot of the project, identifying areas of concern, performing some simple reasoning etc. Most importantly, the communication channels need not be changed for this.

2. BACKGROUND

Processing emails to extract semantic information is not new. Examples include email classification [2], the automatic organisation of emails [1], summarization of email threads [8], conversation detection in emails [3], email indexing [9], as well as various levels of extraction of semantic knowledge from emails, e.g. [6].

The *Infoplaza* project differs from previous work in that it addresses the problem of project management within a fairly well-defined domain. The potential lies in the combination of state-of-the-art techniques and its use in a practical application. Furthermore, it tries to strike the balance between manual and fully automated processing in that all querying and analysis results are linked straight into the documents making it easy to track where certain facts have been derived from (and whether they have actually been extracted correctly).

3. SYSTEM DESCRIPTION

3.1 System Overview

The system is composed of two separate applications. The first one, the *Core Application*, collects email messages from an email server, indexes them and their attachments and extracts their meta-data. Then the emails are processed using NLP components to extract useful knowledge from them. The second one, the *Web Application*, provides a Google-style interface to enable semantic search over the extracted knowledge. It provides access to (automatically extracted) key facts, people, organizations, documents.

3.2 Email Processing with UIMA

The key part in the *Core Application* is integrating NLP tools within the Unstructured Information Management Architecture (UIMA)[1] framework. Each email message col-

[1] `http://incubator.apache.org/uima/`

SemSearch 2009 April 21st, 2009, Madrid, Spain

lected is passed into a pipeline of UIMA analysis engines after encoding their textual content and meta-data into the UIMA's CAS (Common Annotation Scheme) by defining a new UIMA type system for emails.

The CAS representing the email message will hold annotations performed by the various analysis engines. The pipeline consists of a message pre-processor to filter out (annotate) unwanted contents such as quoted texts and signatures using simple heuristics. Then an aggregate *OpenNLP*[2] parser is used for tokenization, part-of speech tagging and full parsing as well as named entity recognition. These annotations feed into the knowledge extraction phase.

3.3 Knowledge Extraction

The annotated version of the content of the emails and their meta-data is fed into UIMA's CAS consumers to extract a knowledge base.

Email meta-data is a very straightforward type of knowledge that is directly extracted from the various email headers.

Another sort of knowledge is extracted from the actual email contents in form of triples, e.g. subject-verb-object relations as used in question-answering systems [5]. Annotations produced as described above are used to generate these triples. The triples can be used then to query the database and to extract relations that exist between different named entities (e.g. "find me everything that Dyaa has ordered").

Automatically extracted data can be complemented by ontological knowledge developed for a specific application domain.

A sentiment analysis step identifies emails of high importance. The classifier we use in the prototype (using training data from the internet movie database) distinguishes between "negative" and "positive" emails. Negative emails are considered important and are flagged up for the project leaders.

3.4 Architecture

Figure 1 sketches the main logical components in the *Infoplaza* architecture. Due to space restrictions we want to pick out a few points only.

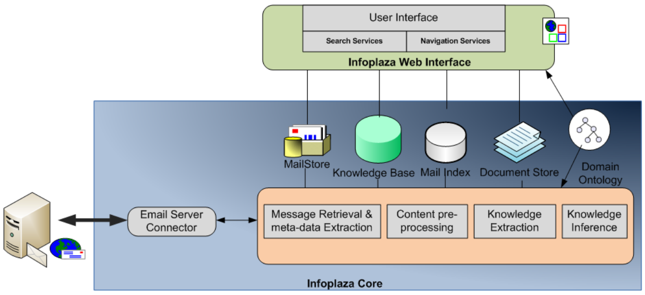

Figure 1: ***Infoplaza*** **architecture.**

Resources in the *knowledge base* are repesented as an RDF graph. The knowledge base has plug-in architecture and the initial implementation uses the Jena[3] Semantic Web framework backed with a MySQL database for persistence. SPARQL queries over the RDF graph where used to navigate the data.

The *web interface* was built using Clojure[4] and the Compojure framework on top of Jetty. It allowed rapid prototyping and java integration. The web interface was built using RESTful services which allows integration with other applications.

4. CONCLUSIONS

The paper presents a prototype of a project management platform based on the UIMA framework. Other than most academic papers it does not aim at significantly advancing the state of the art but instead it demonstrates how existing NLP and knowledge extraction techniques can be employed in a practical and scalable system. We will demonstrate the system using two document collections: the email corpus collected as part of this project as well as the Enron corpus.

Acknowledgments

The project is a partnership between Active Web Solutions (AWS)[5] and the University of Essex and funded by a grant from the Technology Strategy Board under the Knowledge Transfer Partnership scheme, grant number KTP006743, as well as the R&D division of AWS.

5. REFERENCES

[1] Cselle, G. Organizing email. Master's thesis, ETH Zurich, 2006.

[2] de Carvalho, V. R., and Cohen, W. W. On the collective classification of email "speech acts". In *Proceedings of SIGIR* (2005), pp. 345–352.

[3] Erera, S., and Carmel, D. Conversation detection in email systems. In *Proceedings of ECIR* (2008), pp. 498–505.

[4] Jurafsky, D., and Martin, J. H. *Speech and Language Processing- 2^{nd} edition*. Prentice-Hall, 2008.

[5] Katz, B., Lin, J., and Felshin, S. Gathering Knowledge for a Question Answering System from Heterogeneous Information Sources. In *Proceedings of the ACL/EACL Workshop on Human Language Technology and Knowledge Management* (Toulouse, 2001).

[6] Scerri, S., Handschuh, S., and Decker, S. Semantic email as a communication medium for the social semantic desktop. In *ESWC* (2008), pp. 124–138.

[7] Siorpaes, K., and Hepp, M. Games with a purpose for the semantic web. *IEEE Intelligent Systems 23*, 3 (2008), 50–60.

[8] Wan, S., and McKeown, K. Generating overview summaries of ongoing email thread discussions. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics* (2004), p. 549.

[9] Wu, Y., and Oard, D. W. Indexing emails and email threads for retrieval. In *Proceedings of SIGIR* (2005), pp. 665–666.

[2] http://opennlp.sourceforge.net/

[3] http://jena.sourceforge.net/

[4] http://clojure.org/

[5] http://aws.net