Chapter 11
Enhancing Information Extraction with Context and Inference:
The ODIX Platform

Hisham Assal
California Polytechnic State University, USA

Kym Pohl
California Polytechnic State University, USA

Franz Kurfess
California Polytechnic State University, USA

Emily Schwarz
California Polytechnic State University, USA

John Seng
California Polytechnic State University, USA

ABSTRACT

Natural Language Processing (NLP) provides tools to extract explicitly stated information from text documents. These tools include Named Entity Recognition (NER) and Parts-Of-Speech (POS). The extracted information represents discrete entities in the text and some relationships that may exist among them. To perform intelligent analysis on the extracted information a context has to exist in which this information is placed. The context provides an environment to link information that is extracted from multiple documents and offers a big picture of the domain. Analysis can then be provided by adding inference capabilities to the environment. The ODIX platform provides an environment for bringing together information extraction, ontology, and intelligent analysis. The platform design relies on existing NLP tools to provide the information extraction capabilities. It also utilizes a Web crawler to collect text documents from the Web. The context is provided by a domain ontology that is loaded at run time. The ontology offers limited inference capabilities and external intelligent agents offer more advanced reasoning capabilities. User involvement is key to the success of the analysis process. At every step of the process, the user has the opportunity to direct the system, set selection criteria, correct errors, or add additional information.

DOI: 10.4018/978-1-4666-2494-8.ch011

INTRODUCTION

Over the last decade, advances in search engine technology have made them indispensable tools for many users, be it in their private or professional lives. Their benefits are so obvious and far-reaching that it becomes very easy to overlook some of the limitations that they still have. In this chapter, we investigate the needs of analysts or knowledge workers when trying to find pieces of information and knowledge in distributed document repositories, the most widely available one being the World Wide Web. Our initial target area is intelligence analysis, where an analyst is faced with a mountain of documents that may contain possibly relevant information for a particular problem or domain they are investigating. Our assumption is that the analyst has an initial conceptual or formal model of the domain, reflecting the knowledge they have already acquired. Their task now may have two major components: one, to enhance the domain model by incorporating additional knowledge as they investigate the new documents, and two, to identify interesting knowledge items that may be relevant for a particular purpose. An analyst tasked with examining terrorism financing, for example, will have access to existing and historical investigations of this problem, which are used to form the initial conceptual model. Gaining access to additional sources of information, be it public ones from the World Wide Web, or proprietary ones from intelligence agencies or other organizations, allows the agent to expand the model. In the larger context of dealing with terrorism in general, another task of the agent may be to identify suspicious activities, for example the planning of a specific terrorist activity such as a bombing.

A traditional approach to this problem consists of analysts reading documents, bookmarking and annotating them, and entering important pieces of information into databases. These databases can then be queried during the search for relevant information. If the analyst encounters information that may indicate an ongoing plot, an alert could be raised to initiate further investigations. This scenario clearly does not use analysts nor computer systems very effectively, although it apparently was applied by intelligence services for the analysis of the Afghan Wikileaks documents released in the summer of 2010 (Wikileaks, 2010; Schmitt, 2010). The most obvious problem is the amount of time required to read all documents in the repository. This may eventually be necessary in a situation like the Wikileaks documents, but is not practical with a virtually limitless repository such as the World Wide Web. Using the Web as a source requires assistance from a search engine, combined with browsing to follow links on Web pages. Search engines present their results typically as an ordered list of links to documents that are judged to be relevant by their ranking algorithms, again requiring the analyst to read documents, extract information, and enter it in a database. Another limiting factor is the use of databases, which rely on relatively fixed structures for the entries. While it is possible to use databases to store knowledge with a complex internal structure, this structure should be exposed to the user without having to use traditional database methods like database query languages.

Another important aspect where computer support can be essential is the capability to bookmark and cross-reference documents. This can be done with relatively simple means like hyperlinks, but it becomes quickly impractical to do so manually. It is also possible to use search engine technology and create a (reverse) index that lists all occurrences of words (or, to be precise, strings) in the documents. This can be the basis for a Web of metadata linking together entities (e.g. persons, locations) in the documents. The use of natural language processing methods enables a limited degree of named entity recognition, plus the extraction of relationships between entities. This results in a Web of metadata that overlays the original set of documents, possibly incorporating additional knowledge as well. Such an approach