# Web Service Architectures for Text Mining:
## An Exploration of the Issues via an E-Science Demonstrator

*Neil Davis, The University of Sheffield, UK*
*George Demetriou, The University of Sheffield, UK*
*Robert Gaizauskas, The University of Sheffield, UK*
*Yikun Guo, The University of Sheffield, UK*
*Ian Roberts, The University of Sheffield, UK*

**ABSTRACT**

*Text mining technology can be used to assist in finding relevant or novel information in large volumes of unstructured data, such as that which is increasingly available in the electronic scientific literature. However, publishers are not text mining specialists, nor typically are the end-user scientists who consume their products. This situation suggests a web services based solution, where text mining specialists process the literature obtained from publishers and make their results available to remote consumers (research scientists). In this paper we discuss the integration of web services and text mining within the domain of scientific publishing and explore the strengths and weaknesses of three generic architectural designs for delivering text mining web services. We argue for the superiority of one of these and demonstrate its viability by reference to an application designed to provide access to the results of text mining over the PubMed database of scientific abstracts.*

*Keywords: bioinformatics; distributed applications; text mining; text processing; Web services*

## INTRODUCTION

With the explosion of scientific publications it has become increasingly difficult for researchers to keep abreast of advances in their own field, let alone trying to comprehend advances in related fields. Due to this rapid increase in the quantity of available electronic textual data both by publishers and third party providers, automatic text mining is of increasing interest to extract and collate information in order to make the scientific researcher's job easier. Some publishers are already beginning to make textual data available via Web services and this trend seems likely to increase as new

uses for data provided in this manner are discovered. Not only does the internet provide a means to accelerate the publishing cycle, it also offers opportunities for new services to be provided to readers, such as search and content-based information access over huge text collections.

It is not envisioned that publishers themselves will provide technically complex text mining functionality, but that such functionality will be supplied by specialist text processors via "value added" services layered on top of the basic Web services supplied by the publishers. These specialist text processors will need domain expertise in the scientific area for which they are producing text mining applications. However they are unlikely to be the research scientists using the information, because of the specialised knowledge required to build text mining applications. Starting with the presumption of three interacting entities: publishers, text mining application providers and consumers of published material and text mining results, we discuss in this paper a variety of architectural designs for delivering text mining using Web services and describe a prototype application based on one of them. In the rest of this section we review some of the context and related work pertaining to this project.

## Text Mining

Text Mining is a term, which is currently being used to mean various things by various people. In its broadest sense it may be used to refer to any process of revealing information, regularities, patterns or trends, in textual data. Text Mining can be seen as an umbrella term covering a number of established research areas such as information extraction (IE), information retrieval (IR), natural language processing (NLP), knowledge discovery from databases (KDD), and so on. In a narrower sense it requires the discovery of new information, not just the provision of access to information existing already in a text or to vague trends in text (Hearst, 1999). In the context of this paper, we shall use the term in its broadest sense. We believe that, while the end goal may be the discovery of new information from text, the provision of services which accomplish more modest tasks are essential components for more sophisticated systems. These components are therefore part of the text mining enterprise, and lend themselves more freely to being used in Web services architecture.

Text mining is particularly relevant to bioinformatics applications, where the explosive growth of the biomedical literature over the last few years has made the process of searching for information in this literature an increasingly difficult task for biologists. For example the 2004 baseline release of Medline contains 12,421,396 abstracts, published between the years of 1902 and 2004, of which 4,391,392 (around 35 percent) were published between 1994 and 2004.

Depending on the complexity of the task, text mining systems may have to employ a range of text processing techniques, from simple information retrieval to sophisticated natural language analysis, or any combination of these techniques. Text mining systems tend to be constructed from pipelines of components, such as tokenisers, lemmatisers, part-of-speech taggers, parsers, n-gram analysers, and so on. New applications may require modification of one or more of these components, or the addition of new bespoke components; however different applications can often re-use existing components. The exploration of the potential of text mining systems has so far been hindered by non-standardised data representations, the diversity of processing resources across different platforms at different sites and the fact that linguistic expertise for developing or integrating natural language processing components is still not widely available. All this suggests that, in the current era of information sharing across networks, an approach based on Web services may be better suited to rapid system development and deployment.

## Web Services

The World Wide Web Consortium (W3C) defines Web services as "a software system designed to support interoperable machine-to-

# A Spanning Tree Based Approach to Identifying Web Services

Hemant Jain, Huimin Zhao and Nageswara R. Chinta (2004). *International Journal of Web Services Research (pp. 1-20).*
www.igi-global.com/article/spanning-tree-based-approach-identifying/3034?camid=4v1a