



## Text and Concept mining briefing document

Andrew N. Edmonds  
 Scientio, LLC.  
 Friday, 19 October 2007  
 Version 1

<http://www.scientio.com>  
[support@scientio.com](mailto:support@scientio.com)

### Contents

Introduction .....	1
Basics of text and concept mining .....	2
Applications .....	2
Indexing – what other document is like this document? .....	2
Categorisation – What’s the topic of this piece of text out of this set of possible topics? .....	2
Auto summarisation – which parts of this text or set of texts best summarize the whole? .....	3
Natural language conversations – talking to and advising your customers automatically .....	3

### Introduction

This document is intended to briefly cover some of the application areas in text and concept mining that Scientio products are useful for.

Scientio has several software products in this area that can be very useful in solving problems involving semi-structured and unstructured data. In effect semi-structured data means XML or near neighbours like HTML, and unstructured data means text.

These products are already in use in demanding applications worldwide.

Rather than confuse you by detailing the exact function of each product, which you can find described in more detail on our web site <http://www.scientio.com>, we’ll concentrate here on the kinds of problems that can be solved, and leave the details of whether it’s product X you need, or a combination of X and Y, to further discussions.

## Basics of text and concept mining

Text books on the subject of text mining will tell you there are effectively two approaches to processing text with computers in order to extract some kind of meaning from it.

The first uses word frequencies and statistics based on them to generate analyses, and the second, known as natural language processing (NLP), tries to parse text into its grammatical elements in order to extract meaning.

Scientio has opened up a third form of processing. We call this *concept mining*. This uses a model of English or the language of the texts if not English, to extract the likely concepts implied in the text. Concepts have structures that words don't. Concepts are organized into a small number of tree shaped structures, and Scientio has invented new algorithms to take advantage of these structures that allow users to do new things with text, and do old things more effectively. Scientio also has a good implementation of a conventional text mining algorithm based on word frequencies, and is also developing an NLP product.

## Applications

### Indexing – what other document is like this document?

A search engine, such as Google, indexes individual keywords across a wide range of documents. These have become very efficient if the number of keywords is small.

However, there are many kinds of questions a search engine can't answer. Over a year ago a customer asked us to help him to find duplicate or near duplicate documents in a collection his company had of over 1 million documents. It surprised both of us to find that there were no commercial solutions to this problem. Using our ConceptMine product we were able to create a numeric signature, much like a map reference for each document, and supply software that would enable him to keep track of the signatures, and find the closest ones to any given document really efficiently. This is currently *the fastest* algorithm for finding similar documents, since the processing time required grows only with the log of the number of documents, whereas the best competitor's grows linearly.

Our customer has had this product in production for 18 months and is currently handling more than 4 million documents. – It also runs in Spanish.

Obvious applications of this technology are plagiarism detection, near duplicate detection and search engines that cluster documents by content.

### Categorisation – What's the topic of this piece of text out of this set of possible topics?

This is a conventional text mining kind of task. Text mining has been used for some time to try to route and store documents automatically based on their content.

One interesting variation on this came from a customer just two months ago. The customer was interested in mining the sentiment expressed in blogs about various trade names such as ‘Coca Cola’ or ‘David Beckham’. He wanted to analyse the text in weblogs that mentioned these trade names to work out if they were positive or negative about the trade name. This customer currently takes in a third of the blog postings created in English, and intends to take in 75%. Using our product XML Miner we were able to train our text mining algorithm to recognise these postings with around 80% accuracy. We created a web service to process text sent by the customer and respond with the inferred sentiment, and the customer has built a new web site around this showing ‘swingometers’ on a wide range of trade names. At the time of writing this has passed trials and is about to go live.

Interesting variations on this are things like automatically handling and answering emails – one large call centre customer told us that 25% of their traffic now comes from emails.

### **Auto summarisation – which parts of this text or set of texts best summarize the whole?**

Sometimes customers have a large collection of texts that they want to summarize. These could be feedback from their customers, suggestions, etc. Although we could use the categorization technique used above to generate statistics on, say, how negative the comments were, our customers are more likely to be interested in a qualitative, as opposed to quantitative analysis. For instance, if a company has asked its employees to suggest ways to improve business, it would be pointless to discard the least frequent suggestions as they might be the most valuable. Rather, a company would be interested in a response like: ‘out of 2000 suggestions there were 10 discrete ideas, and here they are’.

We are currently working with an external company to develop a general technique to analyse and summarize customer suggestions.

### **Natural language conversations – talking to and advising your customers automatically**

A conversation is the most natural ‘user interface’ that we could create. Many users who are not particularly computerate can nonetheless cope with SMS messages and emails, and that kind of textual conversation. Chatbots are pieces of fairly simple artificial intelligence that can simulate a conversation. They are currently mostly for entertainment, though they have some use on websites – for instance most banking sites seem to have one. Chatbots essentially follow a script, which means that every possible input and response needs to be coded. A limited amount of artificial intelligence is provided to add flexibility to the scripting. The current state of the art open source chatbot contains a script with 14,000 rules. The vast bulk of these rules are concerned with handling the many different ways the same question can be asked. We have applied our concept mining capabilities to Chatbots in two ways. If we analyze the concepts in the user’s inputs, rather than the words, we can reduce the script size dramatically. By doing this we have reduced the number of rules to 3,000. Secondly, the concept structures inherent in language can be used to generate responses. For instance, if you type ‘what is a horse?’ into the industry standard chatbot it responds with a long answer. Ask it ‘what is a pony?’ and it has no response. This is because it doesn’t know that a pony is a kind of horse – if it knew this it could

respond with the same answer. A very large and complete network of this ‘is a kind of ‘data is already built in to our concept mining tools, and we have used this to extend the range of responses created by our new Chatbot called ScientioBot. This is provided as a web service, so that users can create their own front ends for SMS, Email, messenger and web sites.

A separate document talks about rule based systems, but they overlap here. Scientio is generating a questionnaire front end for its rule based systems that takes a rule set and asks questions that will enable it to satisfy the requirements and create an output. If the rule set represents the rules for obtaining housing benefit, for instance, then the questionnaire software will pick the most salient questions to answer first and continue to ask questions, ignoring rules that are no longer relevant because of previous entries, until a response can be generated. This software can produce ten questions at a time or just one question. The latter is, of course, a conversation. Scientio is combining these two technologies into a single product, so users will be able to hold a (limited) conversation with the product before being lead into a questionnaire type session.

Applications for this are governmental websites, legal and ‘compliance’ advice, automated fault finding and support and many more.

Current technology requires a large amount of specialist time to generate the scripts. This is the main barrier to take up. Scientio’s approach dramatically reduces this effort, and non technical subject specialists can be used.

Scientio are currently developing this product with a specialist legal advice company.