The mass of documents is increasing at an alarming rate, and it’s very likely that this rule applies to your organization as well. The vast amounts of reports, emails, notes and other documents created daily by yourself and your organization add ever increasingly to this burden. Ideally, every document should from its inception be annotated with relevant, coherent and complete metadata. The searching of documents would henceforth be facilitated and more enjoyable. But how do we deal with a set of documents that, at the outset, have not been annotated and categorized?

The experts in natural language processing and data mining of the DETI group have developed an approach to aid you in discovering relevant information concealed within your mass of documents: we refer to it as conceptual navigation.

CONCEPTUAL NAVIGATION IS DIVIDED INTO 4 STEPS:

1. Extraction of terms from your document mass
2. Similarity calculation between terms and graphical visualization of related terms
3. Searching of documents containing specific terms
4. A concise view of the documents’ contents

Figure 1 shows these 4 steps incorporated in a user interface. Our term extractor probes your document mass in order to extract the most relevant terms (FIGURE 1, TOP-LEFT WINDOW). Our approach is based on well established statistical methods of computational linguistics.
Our module of related word search is also built upon statistical methods for analyzing your documents and measuring which terms are similar in your context. The resulting lists are to be contrasted with static lists of synonyms found in dictionaries. Our approach is based on a similarity analysis that is a function of your domain and documents. Conceptual navigation then becomes possible through a graphical representation of the related terms (FIGURE 1, TOP-RIGHT WINDOW). You can navigate from one term to the next with a simple click.

You notice a term of particular interest and are curious to see which documents contain that term? A simple click on the word in the graphical user interface triggers a display of the corresponding list of documents (FIGURE 1, BOTTOM-LEFT WINDOW).

Since document titles are not often meaningful, a glimpse of the document is given by listing some of the relevant terms contained within. (FIGURE 1, BOTTOM-RIGHT WINDOW).

The prototype has been developed for the purpose of highlighting various technologies mastered by our experts. Of course, this prototype can be further extended, and in several directions to meet your specific needs.

- You already have a list of terms? That list can replace the result of the extractor.
- You already have a list of synonyms? That list can be added to our similarity computation results.
- You find that the glimpse provided by the keyword list is too limited in scope? We can display some of the document’s phrases using automatic summarization techniques.
- You seek a more refined information search with multiple keywords? We can integrate more advanced technologies in information retrieval.

RESEARCH AND DEVELOPMENT:
Jean-François Lavallée, M.Sc., research agent, DETI team, CRIM
Caroline Barrière, Ph.D., researcher, DETI group, CRIM

INFORMATION:
Sacha Leprêtre, Team Director, DETI, CRIM
Tel.: 514 840-1238 or 1 877 840-2746 - sacha.lepretre@crim.ca

www.crim.ca