DETECTION OF QUASI-DUPLICATES TO REDUCE YOUR MASS OF DOCUMENTS

In our information age, it proves challenging to maintain a singular version of a given document despite the best of our intentions and sound work ethics. The customary practices of storing a document for backup purposes, of exchanging it through email, of preserving a manuscript version of it, are all contributing factors to information overload associated with the document’s content. The information overload introduces extraneous artifacts that may impede the performance of document management systems.

The experts in natural language processing of the DETI group apply a methodology based on the Jaccard similarity to a set of documents, in order to identify the subsets of documents that share a common base and whose contents have been slightly modified.

THE ADVANTAGES OF THIS METHOD ARE:

1. It is not influenced by alterations pertaining to page layouts or to the order of content presentation.

2. It requires no preliminary treatment by a version control system, and is directly applicable to your file directories.

In a previous study, this method achieved a reduction of the mass of documents of 37%.

1 LAVALLÉE, Jean-François, BARRIÈRE, Caroline, Analyse textuelle dans le domaine du recrutement, IC2011, volume 2, numéro 1, automne 2011 (To be published).
Group together largely similar documents (FIGURE 1) in order to preserve the most relevant instance and/or to apply the same system management approach to these documents.

Get informed about the possibilities of integrating to your document management system this quasi-duplicate detection technology developed by CRIM’s Development and Internet Technologies team (DETI).

RESEARCH AND DEVELOPMENT:
Jean-François Lavallée, M.Sc., research agent, DETI team, CRIM

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

www.crim.ca