

CAT

CONTENT ANALYSIS TOOLKIT

Unstructured information in the corporate environment comes in different sizes and shapes, e.g. e-mail, reports, research papers, electronic manuals, web pages, to name a few. Many organisations have different document repositories containing relevant information on a variety of topics. These repositories are usually very large - consisting of thousands of documents and millions of words. The question is: how can one use these documents to assist in accomplishing a certain task without literally having to read all these documents to find relevant information? Furthermore, how can one get an overview of the contents of a collection of documents without opening each document individually?

Indutech's Content Analysis Toolkit was created with just this idea in mind - a tool to assist users to analyse great volumes of information - in the form of electronic documents - to arrive at relevant content quickly and easily. Suppose you found a document with good information on a specific topic; further suppose you want to find another document having the same essence. Imagine being able to hit the "Find me another document like this..." button and ending up with other documents that are highly relevant to the topic you had in mind.

How often does one sit with a list of twenty documents, wishing "if only I knew what these documents are about without having to even open them". Envisage clicking on a document and getting a list of all the highly relevant terms describing the content of the document. Further imagine having to collaborate with an overseas colleague you have never met - wouldn't it be nice if you could get an information interest profile of him/her before your first meeting by analysing a number of documents this person has authored? With CAT a whole new landscape of possibilities is opened in terms of information discovery.

CAT has the following features:

- Provide an overview of the various topics - each topic being described by a number of key terms and associated relevance scores - embedded in large document collections
- Per topic identified, list all documents associated with this topic
- For each topic, show all significantly similar topics identified with their associated documents
- Provide a vocabulary, consisting out of key terms, for a given collection of documents
- For each document, provide the key terms describing the document and the topics related to it
- Present the user with documents that deal with the same topic as a given document
- Automatically group documents into topics based on the content of the documents
- List the key terms describing a certain individual based on the documents associated with him/her
- Detect duplicate or near-duplicate documents irrespective of the difference in filename

The following are some of the popular applications of CAT:

- Assist post-graduate students in identifying which documents to read, and determining the "gaps" in their research documentation
- Get an overview of the topics covered in a body of knowledge (e.g. PMBOK) or a given journal, compile topic-time relations as well as topic-author relations using CAT's output
- Identify "interest profiles" for persons or groups, determine the overlaps between such "interest profiles" and identify "unseen documents" for possible exchange between such persons or groups
- Use the web-based CAT interface as an information portal to access documentation in a topic guided manner