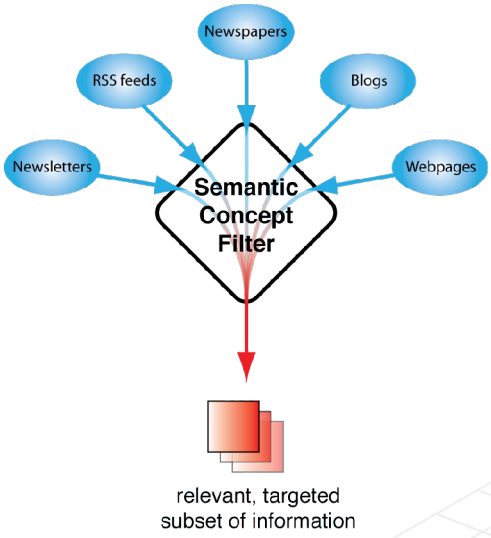
InfoCodex
Semantic Technologies

Phone +41 (0)81 750 53 00
E-mail support@infocodex.com
Web www.infocodex.com

Semantic Concept Filters for Enhanced Media Monitoring



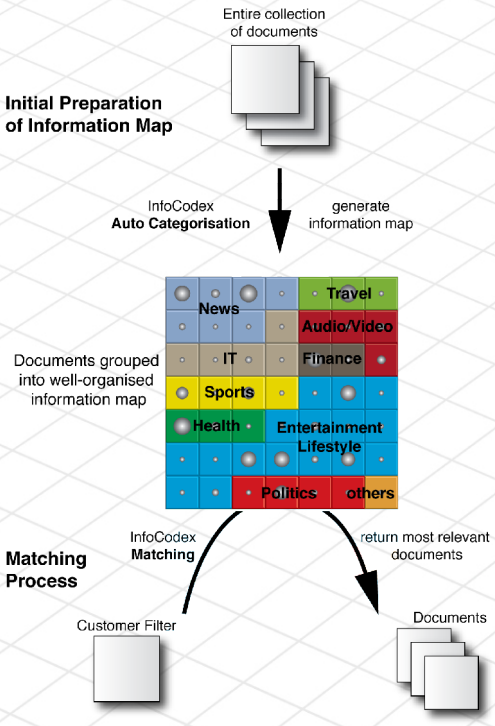
1 Scenario

Media monitoring services have to maintain large pools of documents which are continuously extended by the latest information from newspaper articles, web pages, RSS feeds, newsletters, blogs, and other sources. However, their customers are hardly interested in all the information; they will want a small but qualified subset targeted to their specific needs. How can the relevant information for a specific customer be filtered out and retrieved from the increasing flood of information?

The impact of media intelligence critically depends on the quality of the retrieval as well as on the ranking of the selected documents, i.e. on how well the retrieved documents match the user's current interests and on whether the most relevant ones are topping the list. Finding the best matching documents for a particular customer is a time consuming and resource intense task, as the number of potentially relevant documents can be very large.

The diversity of available documents is enormous, and a matching simply based on the matching of keywords is usually not sufficient. This is further exasperated if the available documents are written in different languages. To maximise impact, an automatic *content-specific* matching of the available information with the customers' needs is required.

A fully automated solution for efficient and high-quality semantic search and match that can be integrated into media monitoring systems would yield substantial benefits.



2 InfoCodex's Semantic Technology for Media Monitoring

InfoCodex can **understand** the thematic content of and **categorise** documents fully automatically (i.e. without human intervention and without any preceding, time-consuming, and case-specific training by means of master text corpuses) – and this across different languages. This is the quintessential point that differentiates InfoCodex from all other products.

InfoCodex employs a scientifically founded similarity scaling (for the comparison of queries or concept filters with documents, and the content of documents among themselves, respectively) which enables an objective ranking of search results as well as a matching of documents.

Using this patented information analysis technology, the documents are automatically categorised according to their content and then organised into an "information map". The structure of the information map is determined automatically by considering the *thematic content* of all available documents and is optimised dynamically as the pool of documents changes.

The customer's needs are stated by means of concept filters (queries) that describe the specific fields of interest. These concept filters can be any text in natural language and are not limited to a simple set of keywords. In order to automate the matching of customer concept filters with available documents,

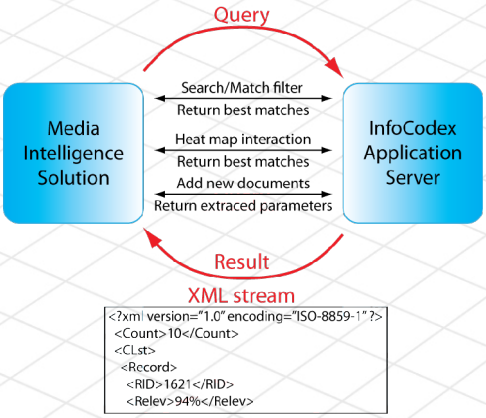
InfoCodex
Semantic Technologies

Phone +41 (0)81 750 53 00
E-mail support@infocodex.com
Web www.infocodex.com

InfoCodex analyses the concept filters in the same way as the available documents, i.e. it analyses their content, and positions them on the information map. InfoCodex's patented **similarity measure** (see below for more details) can now identify the documents which are the best matches to the customer's concept filters, and this in terms of *thematic content* and not only according to common keywords. In simple terms, the documents located on the information map in the same compartment where the customer's concept filter is positioned are the best matches. This process is comparable to an automatic grouping of books (available documents) into a well-organized book-shelf (information map).

This mapping procedure is truly *cross-lingual* and takes into account the *effective content* of the documents; it produces good results even in cases where a simple matching based on keywords is not effective or if the considered documents are written in different languages. For example, an English, French, German, Italian, and Spanish document with equivalent content are recognized by the system as very similar, i.e. have a very high *similarity*, and hence are located at the same position on the information map.

The matching of a specific customer filter with the available documents is done by InfoCodex's *similarity search engine*. The results of a search and match query are not only provided as a list, but can also be displayed in the form of a heatmap to provide enhanced visualisation. This heatmap has the same structure as the infomation map and indicates through colour-coding where in the information map the most relevant hits are (red: most relevant; blue: least relevant). This provides an easy accessible overview, and also enables the targeted finding of relevant documents in specific areas. The heatmap is interactive and allows hits from the selected thematic areas only to be displayed.



3 Integration into Media Monitoring Systems

The InfoCodex search and match engine, including interactive heatmap representations, can be integrated seamlessly into existing media intelligence solutions.

The InfoCodex engine supports several standard protocols including HTTP requests (POST or GET method), Webservices using SOAP, and "Software as a Service" (SaaS). In all cases, the results of the different interactions are returned as an XML stream that can be read and processed by the calling system.

4 Additional Features

A Abstracts and Keywords

The detailed analysis of the retrieved documents by the customer can be very time consuming, even if only the most relevant are considered. The individual documents can be several pages long. In order to provide an easily accessible

Phone +41 (0)81 750 53 00
E-mail support@infocodex.com
Web www.infocodex.com

overview of a document's content and to provide orientation when scouring the result list, InfoCodex fully automatically generates short abstracts and a set of keywords for each document. During the document analysis phase, the sentences which bear the greatest similarity to the essential content of the document (with respect to InfoCodex's established similarity measure), are extracted. This automatically generated abstract is rarely as nicely written as one created by a human author, but it has the advantage that the essential message is concisely worded and restricted to a certain length, which greatly enhances readability.

B Document Families

Very often, the same or very similar documents exist in multiple copies and can be obtained from different sources. For example, identical documents in different formats (PDF, MS Word, HTML, XML etc.), or slightly different or updated versions of a document. In most cases, the customer only wants one copy of such similar documents. Thanks to InfoCodex's ability to understand the content of documents, it can *identify almost identical documents* and combine them into document families. Documents are members of the same family if their contents are close together in the sense of InfoCodex's similarity measure (vector components and descriptors).

C Trend Monitoring

The pool of available documents is constantly expanded by adding the latest information. In many circumstances, the user will want to look at how the available information changes over time, and which thematic areas receive more attention in relation to others. With InfoCodex, such trends can easily be identified and visualised using InfoCodex's heatmaps.