

# A single search portal offering contextual analysis of multimedia news sources



In various domains, analysts have to deal with large amounts of information which is refreshed on a daily basis. News is typically coming from many different open sources and via a number of media, including traditional newspapers, newswires and magazines, internet sites and also television broadcasts. In some use scenarios non-public information sources are equally important. Analysis and monitoring of these news sources is of vital importance for financial analysts, competitive intelligence, patent agencies and national security. Effective analysis of large volumes of both new and archived information requires the availability of text and media mining technology. The Novalist mining tools help to transform a distributed unstructured document collection into a richly annotated multimedia database with flexible options for querying and browsing.

TNO has developed the Novalist system, a multimedia news browser which integrates several media-mining tools. The system facilitates the work of information analysts in the following way: related news stories are clustered, creating threads which are subsequently analysed and annotated with several types of metadata. The annotations make it much easier to find interesting patterns in the data collection and analysts no longer have to search in large sets of unanalysed data, but can search for topical clusters. The Novalist system supports visual browsing of the clusters, along with their extracted headlines, since clusters are visualised in a compact overview window with links to a time axis.

## **Import of multimedia data**

Apart from electronic archives of newspaper documents and magazines, Novalist distills textual representations from audio and video sources, through automatic speech recognition, OCR of text-on-screen, and capture of closed captions and subtitles. All textual documents are stored in an XML database for automated content enrichment.

## **Clustering of related data**

The clustering module of Novalist is based on detection technology, which groups together stories that cover the same topic or event. Topic detection (or topic discovery) deals with dynamic information streams, for which no prior interpretation is available. Clustering is done incrementally, which implies that for a new incoming story, the system has to decide instantaneously to which cluster the story belongs. The clustering technology of Novalist has been tested in the context of an international evaluation event for topic detection systems organized by NIST (National Institute for Standards, USA).

The Novalist system was ranked among the very best scoring systems. Since all the clustering algorithms are unsupervised, no training data is needed.

### Metadata

Novalist makes the hidden semantic similarity between the documents of a cluster explicit in a number of ways. Advanced machine learning and natural language processing techniques are applied to generate concise metadata for each cluster. Headlines for example provide a short indicative text for the contents of the cluster, cluster extracts list the most important excerpts from the contained documents, with redundancy removed. A list of proper names (people, organizations and locations) and important keywords provide more detailed information and can be made available via a mouse click. The clusters found are also automatically annotated with labels from a general thesaurus, providing structured access via a tree structure. All metadata is stored in a relational database, which provides a trivial interface for data mining algorithms.

### Search

Novalist supports various ways of searching the enriched document base: there are browser windows for high-precision search in metadata, search for specific issues via a timeline, and full text search. The individual search modes are interconnected, giving the user maximum control of the search parameters, enabling both recall or precision focused queries.

### Relation networks via Novalink

An additional text mining tool has been developed on top of the Novalist meta-data structure: Novalink. Novalink can dynamically generate a relation network for any person or organization in the database, which can subsequently manually be refined. The relation networks can be exported in RDF form.

### Integration into the workflow of a desk researcher

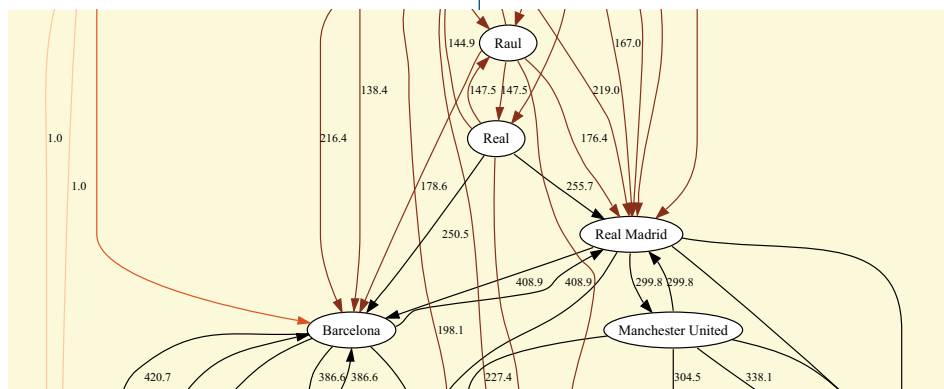
Novalist search results and Novalink networks can be exported to Paranoid, a TNO application developed for the intelligence community. Paranoid enables analysts to extract information for a particular case in a model inspired by (and compatible with) Topic Maps and the Semantic Web. Paranoid offers a wide choice of visualization tools and functionality to assess different hypotheses.

### Added value

Novalist has been evaluated by the Amsterdam-Amstelland police force. The system proved its added value since police investigators were able to find more relevant news items in a database of newspapers, magazines, TV programmes and Web pages using less time.

### Availability

Novalist will be available as product on the basis of a license fee per user/installation. TNO prefers to deliver the product to companies with similar product or tools in their portfolio.



TNO Information and Communication Technology helps companies in many different sectors to become successful innovators. This can result in a new product or service, an improvement, a completely new working method, or a new strategic vision for the future.

Evert van den Akker

Eemsgolaan 3  
P.O. Box 1416  
9701 BK Groningen  
The Netherlands

evert.vandenakker@tno.nl  
www.tno.nl

T +31 50 585 77 12