IBM® Watson Content Analytics Crawler for Online Media

Enables analytics and full text search in online and social media data

Highlights

Search and analyse:

- Data from RSS Feeds
- Social Media Data (Twitter, BoardReader Forum Search Engine)
- Internet web pages (BING Web Search, Google Site Search)

Supports filtering of the crawled articles for relevant content

Fast update crawling

Duplicate detection

IBM Watson Content Analytics is a search and analytics platform that combines the power of content analytics with the scale of enterprise search and includes pre-built integrations for indexing data and content from file shares, databases, collaboration tools and web sites.

IBM Watson Content Analytics Crawler for Online Media extends the existing search and analytics scenarios by data coming either from RSS feeds or from queries against external data sources like BoardReader, BING or Twitter.

Solution Description

The *IBM Watson Content Analytics Crawler for Online Media* can be used to crawl content supplied by RSS feeds. The entries in these feeds are used to download and add the content of the original article to *IBM Watson Content Analytics.*

Furthermore the *IBM Watson Content Analytics Crawler for Online Media* can be used to index data retrieved from third party data providers like BoardReader forum search engine or Twitter to crawl only the relevant social media content out of billions of forum posts, board messages or tweets.

The same applies for crawling Internet web pages. Here the *IBM Watson Content Analytics Crawler for Online Media* uses BING web search as data source to only analyze the web pages that should contain relevant content.

The crawler does not follow any links inside the crawled documents, thus ensuring that only relevant content is added to the analysis.

Unimportant parts of the documents themselves can be filtered out either by using automated algorithms or by specifying regular expressions for filtering.

Technical Information

The online and social media sources to be crawled can be either specified in a configuration file or in a database together with additional fixed data for the crawled documents, e.g. the document language.

The Crawler offers various approaches to ensure a high quality of the crawled documents in the *IBM Content Analytics* index:

- Links inside the crawled documents are not followed, thus ensuring that only relevant content – content listed in the source – is added to the index.
- Added documents are augmented with metadata both from the data source itself (e.g. title, date, author) and optionally also with fixed metadata from the data source configuration (e.g. language).
- By specifying filter patterns or using automated algorithms the administrator can further enhance the quality of document content, e.g. by filtering out header and footer areas.
- Duplicate documents with same content can be detected and ignored.

As the asset stores the URLs of the crawled documents comparable to the standard web crawler, the users of the search and analytics application are able to view the content of the documents as usual with all HTML document inside *IBM Content Analytics*.

Supported Platforms

- *IBM Watson Content Analytics* 3.0 and 3.5
- *IBM Watson Explorer Advanced Edition* 10.0 (Analytical Components)

Supported Formats

- RSS 0.9x, RSS 1.0 / RDF, RSS 2.0, Atom 0.3, Atom 1.0
- BoardReader Forum Search API
- Twitter Search API
- BING Web Search API
- Google Site Search API

For support of non listed versions please contact the Germany Asset Support Center of the ECM Software Services team, reachable via email: **gerasc@de.ibm.com**

Service Offering

- Runtime version per *IBM Watson Content Analytics* system
- Installation and configuration support

IBM Deutschland GmbH
IBM-Allee 1
71139 Ehningen
ibm.com/de

IBM Homepage is reachable below:
ibm.com