IBM® Content Analytics Crawler for RSS Feeds

Ermöglicht Analyse und Volltextsuche in Daten die durch RSS Feeds bereitgestellt werden

Highlights

Suche und Analyse in RSS Feeds

RSS Feeds als URL Liste für das Crawling

Datenquellen wie BoardReader können genutzt werden um die relevanten Inhalte aus Milliarden von Social Media Beiträgen zu crawlen

Relevante Inhalte indizieren, nicht relevanter Inhalt kann herausgefiltert werden

Schnelles Crawlen von Änderungen

Reduntante Artikel können erkannt und ignoriert werden

IBM Content Analytics with Enterprise Search bietet eine skalierbare, sichere und hochwertige unternehmensweite Suche und Inhaltsanalyse. Diese stellt vorgefertigte Integrationen zur Indizierung von Datenbanken, Kollaborationsanwendungen (Blogs, Wikis, etc.) und Inhalte von Dateisystemen zur Verfügung.

Der *IBM® Content Analytics Crawler for RSS Feeds* erweitert die vorhandene Such- und Analyseszenarien um die Suche in Inhalten, die entweder durch RSS (Really Simple Syndication) Feeds zur Verfügung gestellt oder durch Abfragen an Datenquellen wie BoardReader bereitgestellt werden.

Funktionalitäten

Der *IBM Content Analytics Crawler for RSS Feeds* kann verwendet werden um Inhalte zu crawlen die über RSS Feeds zur Verfügung gestellt werden.

RSS Feeds bestehen aus einer Liste von Einträgen mit einer kurzen Zusammenfassung sowie einem Titel, weiteren Metadaten, zum Beispiel ein Link zum Artikel selbst, ein Veröffentlichungsdatum und Informationen über den Autor.

Der *IBM Content Analytics Crawler for RSS Feeds* nutzt diese Informationen um den Artikel herunterzuladen und den Inhalt des Artikels dem Volltextindex von *IBM Content Analytics* hinzuzufügen. Die durchsuchten RSS Feeds können daher als URL Liste für das Crawling betrachtet werden.

Der Crawler folgt dabei nicht den im Artikel enthaltenen weiteren Verweisen, dadurch wird sichergestellt das nur relevanter Inhalt dem Volltextindex hinzugefügt wird, nicht relevante Bestandteile der Seiten können über reguläre Ausdrücke ausgefiltert werden. Darüber hinaus kann der *IBM Content Analytics Crawler for RSS Feeds* genutzt werden um Daten von Anbietern wie „BoardReader Forum Search Engine“ zu indizieren, wodurch sichergestellt werden kann dass aus den Milliarden von Social Media Beiträgen nur die relevanten Beiträge indiziert und analysiert werden.

Technische Informationen

Die zu crawlenden RSS Feeds und Social Media Quellen können entweder in einer flachen Liste von URLs oder in einer Datenbank zur Verfügung gestellt werden, zusammen mit weiteren fest vorgegebene Metadaten für die neuen Dokumente, zum Beispiel die Sprache.

Der Crawler bietet verschiedene Optionen um eine hohe Qualität der Dokumente im *IBM Content Analytics* Index zu erreichen:

- Verweise innerhalb der gecrawlten Dokumente werden nicht verfolgt, dadurch wird nur der relevante Inhalt, der in den Feeds gelistet ist, dem Index hinzugefügt.
- Die Dokumente werden mit Metadaten angereichert die sich sowohl aus den Feed-Einträgen selbst ergeben (z.B. Titel, Datum, Autor) sowie optional auch durch feste Metadaten aus einer Datenbanktabelle (z.B. Sprache).
- Durch die Verwendung von regulären Ausdrücken kann die Qualität der Dokumente weiter erhöht werden, beispielsweise können Kopf- und Fußzeilen ausgefiltert werden.
- Doppelt vorhandene Dokumente mit gleichem Inhalt können erkannt und ignoriert werden.

Da der Crawler sowohl URLs und Inhalt der Dokumente ähnlich dem Standard Web Crawler speichert, können die Benutzer der Such- bzw. Analyseanwendung diese wie gewohnt anzeigen.

Unterstützte Versionen

Aktuell getestete Versionen

- *IBM Content Analytics with Enterprise Search* 3.0
- *IBM OmniFind* 9.*1 Enterprise Edition*
- *IBM Content Analytics* 2.2 (nur für Suchobjektgruppen)

Für Abklärung des Support von abweichenden Versionen kontaktieren sie bitte das Germany Asset Support Center des ECM SWG Services Team unter der E-Mail: **gerasc@de.ibm.com**

Unterstützte Formate

- RSS 0.9x, RSS 1.0 / RDF, RSS 2.0
- Atom 0.3, Atom 1.0

Serviceangebot

- Runtime Version je *IBM Content Analytics with Enterprise Search* System
- Unterstützung bei Installation und Konfiguration

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

Die IBM Homepage erreichen Sie unter:
ibm.com