

Scandio SEBOL Search

Search Engine Based On Lucene

Autor: Christian Koch

Datum: 16. September 2010

Version: 1.02

© Scandio GmbH, München

Inhalt

1. Was ist SEBOL?	3
2. Index-Server	4
2.1. Warteschlange zur Indizierung	4
2.2. Plugin-Abarbeitung	4
2.3. Erweiterte Lucene-Indizierung	4
2.4. Index-Verteilung und –Management.....	5
3. Search Server	5
3.1. Plugin-Abarbeitung	5
3.2. Navigatorensuche	6
3.3. Ergebnissortierung.....	6
3.4. Paging.....	6
4. Client API	6
4.1. Programmiersprachen	6
5. Kontakt	7

© Scandio GmbH

Datei
2010-09SEBOL Kurz.doc

Autor
Christian Koch

Version
1.02

Datum
16. September 2010

Seite
2 von 7

Technische Kurzbeschreibung SEBOL

1. Was ist SEBOL?

Die Scandio GmbH beschäftigt sich seit 2004 mit Suchtechnologien und verfügt über Know-how in der Planung und im Aufbau von individuellen Systemen zur Indizierung und zur Suche von strukturierten und unstrukturierten Daten.

Auf der Basis der Open Source Technologie Apache Lucene entwickelt die Scandio SEBOL (Search Engine Based On Lucene). Scandio SEBOL ist ein in Java geschriebenes Softwaresystem, welches das Indizieren und Suchen in großen Datenmengen vollständig abdeckt und Lösungen zu vielen Problemstellungen anbietet. Scandio SEBOL verfügt über die Möglichkeit, kundenspezifische Plugins über die eigene Pluginschnittstelle einzubinden und ist daher beliebig zu erweitern und hervorragend an den jeweiligen Einsatzbereich anzupassen. Scandio SEBOL kann sowohl strukturierte Datenbanken als auch unstrukturierte Dokumente verarbeiten und ist skalierbar. Die Suchmaschine kann als Standalone-Server oder in einer JSP-Engine betrieben werden (getestet mit Tomcat).

Im Kern besteht Scandio SEBOL aus folgenden Komponenten:

- Index-Server
- Search-Server
- Client – API

Jede Komponente beinhaltet unterschiedliche Module, die die Grundlage des Gesamtsystems bilden. Zur optimalen Skalierbarkeit des Systems sind dies eigenständiger Server, die im Verbund miteinander kommunizieren.

© Scandio GmbH

Datei
2010-09SEBOL Kurz.doc

Autor
Christian Koch

Version
1.02

Datum
16. September 2010

Seite
3 von 7

2. Index-Server

Der SEBOL Index Server enthält alle Module, die zur Erstellung eines Index notwendig sind. Sie können als autarke Server betrieben werden.

Die Kernmodule sind:

- Warteschlangenverwaltung zur Indizierung
- Plugin-Abarbeitung von System- und Kunden-Plugins
- Erweiterte Lucene Indizierung
- Index-Verteilung und -Management

Diese Module nehmen die Indizieranfragen entgegen, führen Vorverarbeitungen durch, speichern diese als Lucene-Index und verteilen diesen Index bei Bedarf auf mehrere Search-Server.

2.1. Warteschlange zur Indizierung

Die Indizier-Queue nimmt die Anfrage per XML-RPC bzw. REST entgegen und legt diesen persistent ab. Sollte der Server gestoppt werden, bleiben die nicht abgearbeiteten Queue-Einträge erhalten und werden bei Neustart des Index-Servers weiter abgearbeitet. Die Warteschlange beinhaltet die Möglichkeit der Priorisierung. Das bedeutet, dass wichtige Index-Anfragen vorgezogen werden können.

Zusätzlich steuert die Warteschlange das Journaling des Index-Servers, welches dafür sorgt, dass alle Aufgaben zwischen dem letzten Backup und einem eventuellen Ausfall des Systems nachvollzogen und erneut abgearbeitet werden können.

2.2. Plugin-Abarbeitung

Sowohl die System-Plugins als auch die kundenspezifischen Plugins werden vor der Indizierung mittels einer frei definierbaren Plugin Chain abgearbeitet. Dabei können die zu indizierenden Daten verändert, erweitert und analysiert werden.

Als System-Plugins sind hier die Content Reader für HTML, PDF und die bekannten Office Formate (mittels TIKA) bereits enthalten. Typische Kunden-Plugins sind die Aufbereitung von Datenbankinhalten für die Indizierung.

2.3. Erweiterte Lucene-Indizierung

Der Schritt der Indizierung baut die Navigatoren (Faceted Search) auf, sofern dies konfiguriert ist. Dabei beherrscht SEBOL nicht nur die klassischen Zählnavigatoren, sondern auch numerische und

© Scandio GmbH

Datei
2010-09SEBOL Kurz.doc

Autor
Christian Koch

Version
1.02

Datum
16. September 2010

Seite
4 von 7

chronologische Bereichsnavigatoren sowie Entfernungen für die Umkreissuche. Navigatoren sind frei definierbar und ihre Funktion ist über Plugins erweiterbar. Final wird mittels des Lucene Index-Writers der tatsächliche Suchindex aufgebaut.

2.4. Index-Verteilung und –Management

Bei hoch frequentierten Seiten steht die Suchgeschwindigkeit im Vordergrund. Daher hat die Scandio einen Sync-Mechanismus entwickelt, der auf Apache Hadoop DFS basiert. Dieser Sync-Mechanismus verteilt im Hintergrund den „neuen“ Index auf die Search-Server und erteilt über ein internes Messaging-System die Freigabe zum Umschalten auf den neuen Index. Weiterhin sind Management Funktionen enthalten, um zum Beispiel die Synchronisation zu pausieren.

3. Search Server

Der Search-Server nimmt die Suchanfragen per XML-RPC oder REST entgegen, führt die Suche durch und liefert das Ergebnis sortiert an den Client zurück.

Um den Lucene-Kern herum wurden dabei wichtige Module hinzugefügt:

- Vor- und Nachgelagerte Plugin-Abarbeitung
- Navigatorensuche (Faceted Search)
- Umkreissuche
- Ergebnissortierung
- Paging

Die Module beinhalten die folgende Funktionalität:

3.1. Plugin-Abarbeitung

Bei der Suche kann es sowohl vor als auch nach der eigentlichen Ausführung der Lucene-Suche kundenspezifische Anpassungen geben. Mittels der SEBOL Plugin-Kette ist es möglich, beide Arten von Plugins zu implementieren. Zum Beispiel wird ein frei konfigurierbares Boosting-Plugin mit ausgeliefert, mit dem ohne Neuindizierung bestimmte Attribute stärker in das Suchranking einbezogen werden können. Die Integration von Benutzerrechten und –rollen ist ebenso über die Plugin-Schnittstelle möglich. Daher kann SEBOL jede Benutzerverwaltung integrieren.

© Scandio GmbH

Datei
2010-09SEBOL Kurz.doc

Autor
Christian Koch

Version
1.02

Datum
16. September 2010

Seite
5 von 7

3.2. Navigatorensuche

Mit jedem Such-Request können Navigatoren angefordert werden. Diese werden in der Regel für die weitere Einschränkung des Ergebnisses verwendet.

3.3. Ergebnissortierung

Grundsätzlich kann das Ergebnis nach jedem Single-Value Attribut sortiert werden. Dies bedeutet, es sind Sortierungen in alphabetischer, numerischer und chronologischer Reihenfolge (bei Date-Attributen) möglich.

Weiterhin wurde die besondere Sortierung nach der Entfernung für die Umkreissuche eingeführt. Diese Sortierung wird nachgelagert durchgeführt, da Entfernung ein relatives Sortierkriterium ist.

Navigatoren können ebenfalls sortiert werden. Hier kommt noch die Sortierung nach der Anzahl (Count) hinzu.

Die Standardsortierung des Ergebnisses erfolgt nach der Relevanz.

3.4. Paging

Beim Paging kann die Startposition und die Anzahl der zu übertragenden Ergebnisse angegeben werden.

4. Client API

Die Basis für das Client API ist das XML-RPC Konstrukt, mit dem der Index-Server und der Search-Server angesprochen werden. Zusätzlich existiert eine Json basierte REST Schnittstelle. Beide Basisschnittstellen haben den identischen Funktionsumfang. Sie beinhalten alle Methoden zum Indizieren und zum Suchen.

4.1. Programmiersprachen

Auf der Basis der XML-RPC Schnittstelle von SEBOL sind derzeit APIs in den Programmiersprachen Java/JSP und PHP implementiert. Beide Klassenbibliotheken haben den identischen Funktionsumfang. Die Entwicklung weiterer Schnittstellen für Programmiersprachen wie zum Beispiel ASP/ASPX, Ruby oder Python ist problemlos möglich.

© Scandio GmbH

Datei
2010-09SEBOL Kurz.doc

Autor
Christian Koch

Version
1.02

Datum
16. September 2010

Seite
6 von 7

5. Kontakt

Weitere Informationen erhalten Sie über

Scandio GmbH

Christian Koch
In der Rosenau 6
81829 München

Tel.: +49 89 244 124-44

Mail: christian.koch@scandio.de

Web: <http://www.scandio.de>

© Scandio GmbH

Datei
2010-09SEBOL Kurz.doc

Autor
Christian Koch

Version
1.02

Datum
16. September 2010

Seite
7 von 7