

Outermedia Highlighter Version 1

Datum: 24. Jun 2015
Verfasser: Andreas Koch

Inhaltsverzeichnis

Outermedia Highlighter Version 1	1
1. Problem des Default Solr Highlighters	2
2. Verbesserung	2
3. Voraussetzungen	2
4. Installation des Patches.....	3

1. Problem des Default Solr Highlighters

Der Default Solr-Highlighter ermittelt alle Terme aus einer Query. Diese Terme werden dann im Text hervorgehoben. Terme in Verbindung mit NOT werden nicht hervorgehoben. Allerdings wird auch nicht überprüft, ob ein Term zu einem Treffer im Dokument führte. Dies soll an folgendem Beispiel verdeutlicht werden:

```
QUERY: Haus OR (Hammer AND Nagel)
TEXT: Ein Hammer liegt im Haus.
```

Dies führt zu folgender Hervorhebung: Ein *Hammer* liegt im *Haus*.

Haus hätte jedoch nicht hervorgehoben werden sollen, da es nur in Verbindung mit Nagel zu einem Treffer führt. Das Wort Nagel kommt im Text nicht vor.

2. Verbesserung

Der Outermedia Highlighter beachtet diese logischen Verknüpfungen der Suchterme und hebt nur diejenigen Terme hervor, die an einem Treffer beteiligt waren.

3. Voraussetzungen

Der Outermedia Highlighter ist ein Patch auf Basis der *Solr Version 4.4*. (Aufgrund einiger Realisierungsdetails im Source Code von Solr ist es derzeit nicht möglich, den Outermedia Highlighter, als Plugin zu realisieren.)

Der Outermedia Highlighter funktioniert in folgendem Umfeld:

- Der Outermedia Highlighter funktioniert mit der Solr Version 4.4. Da die Implementierung in Form eines Patches vorliegt, kann keine Aufwärtskompatibilität gewährleistet werden.
- Für den Outermedia Highlighter muss der Parameter *hl.requireFieldMatch* auf true gesetzt werden. Dies kann entweder in der Datei *solr-config.xml* als Default konfiguriert werden, oder dieser Parameter kann bei der Suche als Request-Parameter angegeben werden.
- Der Outermedia Highlighter funktioniert nicht im Zusammenspiel mit dem Fast Vector Highlighter und nicht mit dem Postings Highlighter. Unterstützt wird nur der Highlighter: `org.apache.lucene.search.highlight.Highlighter`
- Der Outermedia Highlighter funktioniert mit dem edismax-Parser und mit dem Outermedia Query Parser.
- Im Feld, das hervorgehoben werden soll, darf der Parameter *omitTermFreqAndPositions* **nicht** auf true gesetzt sein.
- Der Outermedia Highlighter ist langsamer als der Default Solr Highlighter, weil zusätzliche Überprüfungen durchgeführt werden. Im Gegensatz zum Default Solr Highlighter hat hier auch die Länge eine Query einen starken Einfluss auf die Performance. Deshalb ist der Einsatz von Filtern zu empfehlen, wo dies sinnvoll und machbar ist.

4. Installation des Patches

```
#!/bin/bash
# Patches the given solr.war with the given highlighter patch
# creates a backup of the given solr.war (solr.war.orig)
patchedHighlighterJar=/path/to/patch/dir/sma-solr-highlight-5.3-XXX.jar
solrWar=/path/to/solr-war/dir/solr.war

#two jars have to be patched / exchanged
jarsToPatch="WEB-INF/lib/lucene-highlighter-4.4.0.jar WEB-INF/lib/solr-core-4.4.0.jar"

#backup of solr.war
cp $solrWar $solrWar.orig

cd /tmp
mkdir -p extracted
rm -rf extracted/*
unzip -d extracted $solrWar $jarsToPatch
zip -d $solrWar $jarsToPatch

# patchedFiles sind die Java-Klassen, die im sma-solr-highlight-5.3-XXX.jar
vorliegen
patchedFiles=`unzip -l $patchedHighlighterJar |sed -e "s/\\s\\+/ /g"|cut -d" " -f5-|grep "\.class"`

#löschen der patchedFiles
zip -d extracted/WEB-INF/lib/lucene-highlighter-4.4.0.jar $patchedFiles
zip -d extracted/WEB-INF/lib/solr-core-4.4.0.jar $patchedFiles
cp $patchedHighlighterJar extracted/WEB-INF/lib/
cd extracted/
zip -r $solrWar *
cd ..
```

Das Skript verändert ein solr.war so, dass die gepatchten Versionen der solr/lucene-Klassen verwendet wird (auf Basis des sma-solr-highlight-Artefakts). Hierzu werden folgende Schritte ausgeführt:

- Auspacken des solr-Archives (WAR) mit der aktuellen Version, getestet mit Version 4.4
- Ersetzen aller Klassen des Patches in diesen beiden jars:
 - WEB-INF/lib/lucene-highlighter-4.4.0.jar
 - WEB-INF/lib/solr-core-4.4.0.jar
- Erstellen des Solr-Archives (WAR)

Nun muss das gepatchte solr.war gemäß Installationsanleitung im Jetty als Webapp installiert werden.