Rebecca Schwarz

„BACK TO BASIC“: URL-SEGMENTE BILDEN UND EINE ÜBERSICHT ÜBER DIE KLICKVERTEILUNG MIT KNIME ERHALTEN

Jedes Jahr ein neuer Trend, jede Woche ein neues KI-Tool, um mithilfe der Suchmaschinenoptimierung die eigene Website auf ein vermeintlich neues Level zu heben. Doch welche Seitenbereiche zeigen die größten Potenziale und welche Seitentypen konnten im vergangenen Zeitraum kaum bis keine Klicks generieren?
Dieser Artikel steht im Zeichen von „Back to Basic“. Denn zu Beginn einer Website-Optimierung sollte der Fokus erst einmal darauf liegen, sich eine Übersicht über die bestehenden Seiten zu verschaffen. Deshalb wird hier Schritt für Schritt gezeigt, wie eine Website anhand von Mustern nach Seitentemplates und Seitenbereichen unterteilt werden kann. Mithilfe von KNIME kann damit abgeleitet werden, wie viele URLs sich in den Seitenbereichen befinden und beispielsweise welche Verzeichnisse die meisten Klicks erzielen konnten.
Die Erkenntnisse können für einen SEO-Frühjahrsputz ebenso wie für die Planung eines Relaunchs genutzt werden. Denn hier zeigt sich, in welchen Bereichen wirklich Potenziale stecken!

Überblick mit KNIME verschaffen

Auch in diesem Artikel wird wieder einmal beispielhaft gezeigt, wie die Open-Source-Software KNIME verwendet werden kann, um Aufgaben in der Suchmaschinenoptimierung durchzuführen. Diesmal soll der Fokus auf einem grundlegenden Task liegen, der Segmentierung einer Website. Denn häufig ist gar nicht klar, wie viele Verzeichnisse eine Website hat, welches Ziel diese Seitenbereiche verfolgen und welche historisch gewachsenen Seiten überhaupt noch von potenziellen Kunden angeschaut werden.

Die schlechte Nachricht: Jede Website ist anders, weshalb auch die Segmentierung einer Website sehr individuell ist. Es gibt somit nicht immer einen schnellen Weg, sich eine vollständige Übersicht über eine Website zu verschaffen.

Aber die gute Nachricht: Der vorgestellte KNIME-Workflow kann ganz individuell angepasst werden und ist ab Veröffentlichung des Artikels kostenfrei im KNIME-Community-Hub abrufbar.

Warum KNIME?

KNIME ist eine Software für verschiedenste Aufgaben der Datenanalyse. Das Programm verfügt über eine grafische Oberfläche und arbeitet mit sogenannten Knoten, die aneinandergereiht einen Workflow bilden. Jeder Knoten kann mit Daten eine bestimmte Aufgabe durchführen. Zum Beispiel lesen READER-Knoten Daten ins Tool ein, während WRITER-Knoten verarbeitete Daten in eine Datei schreiben und somit aus dem Tool exportiert werden. Die Vorteile des Tools sind vor allem, dass keine Programmierkenntnisse notwendig sind und einmal erstellte Workflows immer wieder mit neuen Daten durchlaufen werden können. Außerdem ist das Tool kostenfrei nutzbar und wird von einer großen Community unterstützt.

DIE AUTORIN

Rebecca Schwarz ist SEO-Consultant bei der get:traction GmbH. Ihr Arbeitsalltag dreht sich um die Konzeption von SEO-Strategien und die Unterstützung von Kunden in der redaktionellen SEO. Um größere Datenmengen effizient zu verarbeiten und um bei wiederkehrenden SEO-Tasks Zeit zu sparen, nutzt sie die Open-Source-Software KNIME.

Start der Analyse

Der nachfolgende Workflow ist so gestaltet, dass er sich individuell anpassen lässt und eine Vielzahl an Anwendungsmöglichkeiten bietet. Zur Ausführung wird Folgendes benötigt:

» Liste aller URLs einer Domain (Idealfall: der internal_all-Export aus dem Screaming Frog)
» Lokale Installation der KNIME-Umgebung
» Zugang zur Google Search Console

Einlesen der Daten

Der erste Schritt in der Arbeit mit KNIME ist immer das Einlesen der Daten ins Tool. Hierzu wird ein READER-Knoten benötigt. Um sich nicht lange mit der Wahl des korrekten READER-Knotens aufzuhalten, kann einfach per Drag-and-drop die entsprechende Datei in die eigene KNIME-Umgebung gezogen werden.

Je nach Dateiformat wird automatisch der passende READER-Knoten ausgewählt. Für den CSV-Datei-Export aus dem Screaming Frog erscheint in der Oberfläche nun der CSV-READER. Im Dialogfenster des Knotens wird nun eine Vorschau des Datensets angezeigt (Abb. 1).

Für das Einlesen des „internal_all"-Exports muss in der Konfiguration im Reiter „Encoding" UFT-8 oder UTF-16 ausgewählt werden, damit alle Werte des Datensets korrekt dargestellt werden. Alle anderen Einstellungen können beibehalten werden. Mit Klick auf „OK" schließt sich der Knoten. Mit Rechtsklick auf den Knoten „Execute" wird der Knoten ausgeführt.

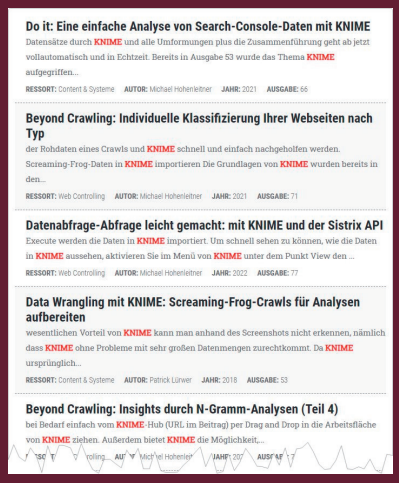
Bei erfolgreicher Ausführung steht das Ampelsymbol unter dem Knoten nun auf Grün. Die Daten sind somit eingelesen und ein weiterer Knoten kann angehängt werden.

Irrelevante Spalten aus dem Datenset filtern

Vor allem in den Crawl-Exporten sind deutlich mehr Informationen enthalten, als für diese Analyse notwendig sind. Um eine bessere Übersicht zu haben, werden deshalb zunächst nicht relevante Spalten aus dem Datenset herausgefiltert. Dies funktioniert mit dem Knoten COLUMN FILTER.

HINWEISE FÜR DEN KNIME EINSTIEG:

Hinweise für KNIME-Anfänger*innen: Die Software KNIME arbeitet mit sogenannten Knoten, die miteinander verbunden einen Workflow ergeben. Zu Beginn werden immer über einen READER-Knoten Daten in die Oberfläche eingelesen. In daran angebundenen Knoten werden die Daten entsprechend verändert, transformiert oder angereichert. Jeder Workflow endet normalerweise mit einem WRITER-Knoten, mit dem die Daten schlussendlich in eine Datei, zum Beispiel eine Excel-Datei, geschrieben werden. Anschließend kann mit dieser Datei dann im entsprechenden Programm ganz leicht weitergearbeitet werden.
Die größten Vorteile der Datensoftware im Allgemeinen sind, dass das Tool kostenfrei verwendbar ist und keine Programmierkenntnisse notwendig sind. Außerdem ist auch für ChatGPT die Software nicht unbekannt und kann bei Problemen mit der Konfiguration helfen. Auf knime.com gibt es umfassende Anleitungen und Dokumentationen zur Einführung ins Tool. Passende SEO-Anwendungsfälle waren und sind außerdem Bestandteil vieler Ausgaben der Website Boosting seit Ausgabe 53 und unter einfach.st/alleknime online abrufbar.



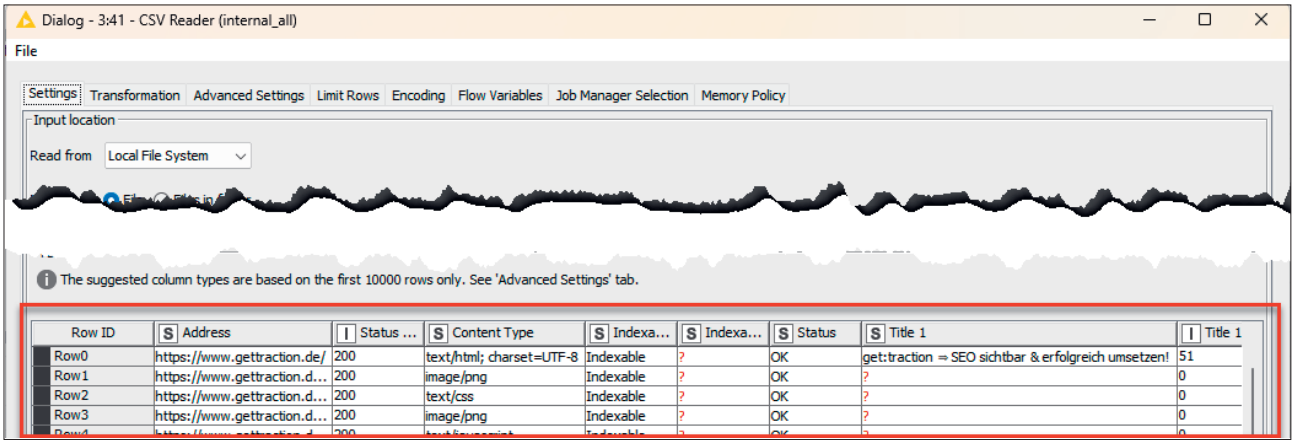


Abb. 1: Konfiguration im CSV-READER

HINWEIS

Leuchtet das Ampelsymbol unterhalb eines Knotens gelb mit einem Warnsignal, kann der Knoten nicht ausgeführt werden. Hierbei kann man einfach mit dem Mauszeiger über das Warnsymbol fahren, um die Fehlermeldung angezeigt zu bekommen. Ein häufiger Fehler ist, dass das Datenset leer ist, weil die Konfiguration eines Knotens dazu geführt hat, dass alle Daten aus dem Datenset gefiltert wurden.

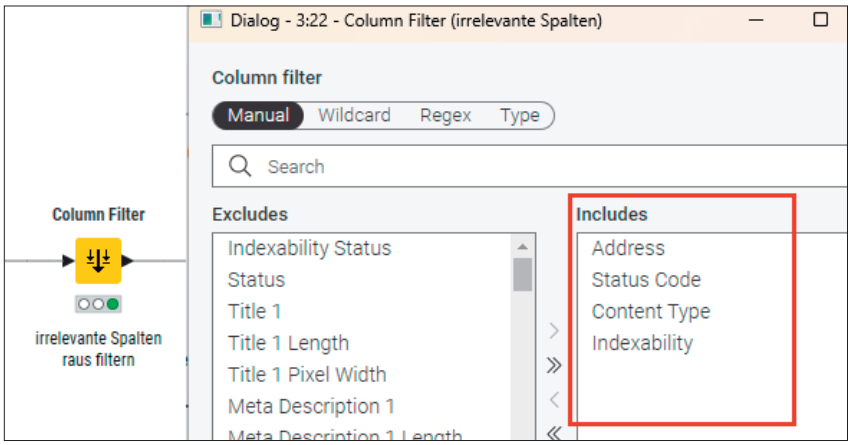
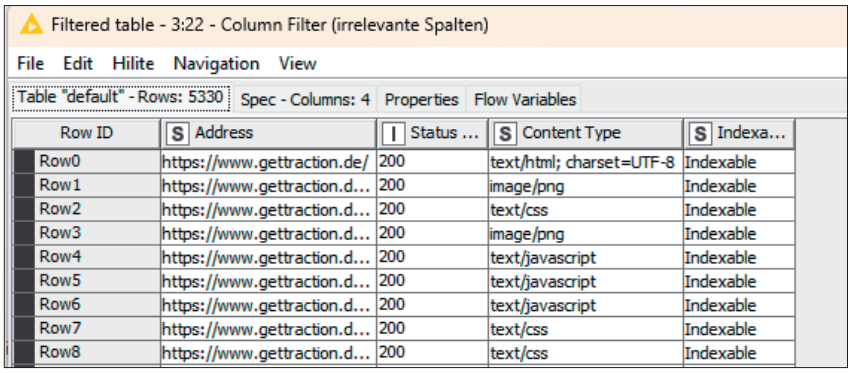


Abb. 2: Konfiguration im COLUMN FILTER



Row ID	Address	Status ...	Content Type	Indexa...
Row0	https://www.gettraction.de/	200	text/html; charset=UTF-8	Indexable
Row1	https://www.gettraction.d...	200	image/png	Indexable
Row2	https://www.gettraction.d...	200	text/css	Indexable
Row3	https://www.gettraction.d...	200	image/png	Indexable
Row4	https://www.gettraction.d...	200	text/javascript	Indexable
Row5	https://www.gettraction.d...	200	text/javascript	Indexable
Row6	https://www.gettraction.d...	200	text/javascript	Indexable
Row7	https://www.gettraction.d...	200	text/css	Indexable
Row8	https://www.gettraction.d...	200	text/css	Indexable

Abb. 3: Vorschau des Datensets nach der Spaltenfilterung

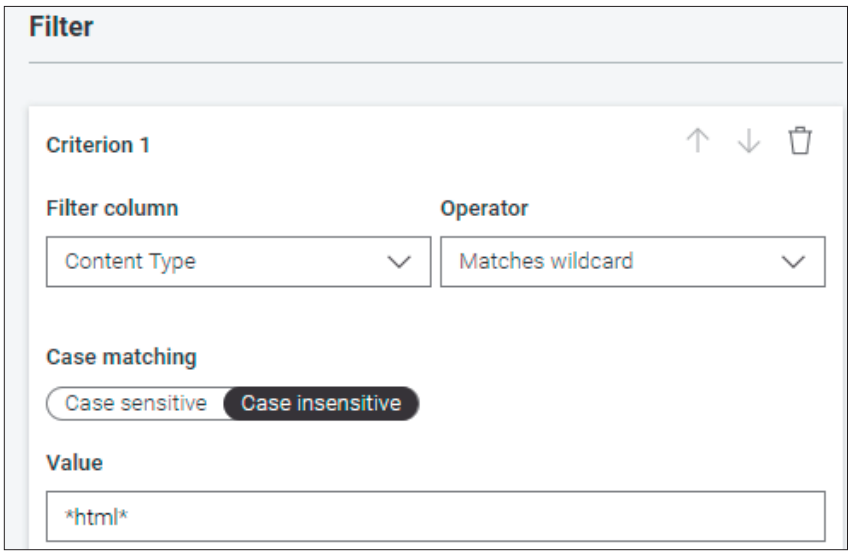


Abb. 3: Vorschau des Datensets nach der Spaltenfilterung

Im Dialogfenster werden dazu nur die wichtigen Spalten in den Includes (rechtes Fenster) beibehalten (Abb. 2).

Mit „OK“ und Rechtsklick auf „Execute“ besteht das Datenset nun nur noch aus den vier ausgewählten Spalten (Abb. 3).

In der hier vorgestellten Analyse sollen nur die HTML-Dokumente des Crawls segmentiert und analysiert werden, weshalb alle anderen Content-Typen noch aus dem Datenset herausgefiltert werden. Anders als im vorherigen Knoten sollen nun Zeilen statt Spalten gefiltert werden, weshalb als Knoten ein ROW FILTER gebraucht wird.

Im ROW FILTER wird als Filter column „Content Type“ ausgewählt. Als Operator wird „Matches Wildcard“ eingestellt und in den Value wird „*html*“ eingegeben. Wichtig ist nun noch, dass als Case matching „Case insensitive“ gewählt wird. Somit wird sichergestellt, dass alle URLs gematcht werden, die als Content-Typen ein HTML-Dokument sind (Abb. 4).

Nun sind die Crawl-Daten ausreichend aufgeräumt und bereit für die Analyse.

GSC-Daten ans Datenset anhängen

Weil bisher nur Crawl-Daten im Tool vorhanden sind und diese noch keine Klicks und Impressionen laut der Google Search Console (kurz: GSC) beinhalten, werden diese Daten nun in KNIME importiert.

Bisher gab es hierfür nur eher komplizierte Wege, um die GSC-Daten via KNIME abzufragen. Ab sofort gibt es hierfür einen viel einfacheren Weg, nämlich ein eigenes Knotenset mit dem Namen SEARCH ANALYTICS. Alles, was dazu notwendig ist und wie die Knoten korrekt konfiguriert werden, kann im Artikel von Mario Fischer in dieser Ausgabe nachgelesen werden.

Verknüpfung der Daten

Sobald die Daten bereitstehen, werden die GSC-Daten mit den Crawl-Daten verknüpft. Wie immer wird dazu auf

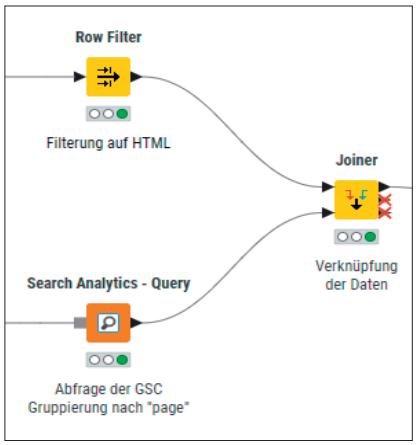


Abb. 5: Verbindung mit dem JOINER

HINWEIS

Hinweis zum JOINER: Diese Art der Verknüpfung wird verwendet, damit alle Daten übernommen werden. Denn es ist in jedem Fall möglich, dass sich in den GSC-Daten URLs befinden, die nicht mehr im Crawl vorhanden sind. Andersherum werden auch in den Crawl-Daten neue URLs existieren, die der GSC noch unbekannt sind. Würde in den Einstellungen nur „matching rows" gewählt werden, würden nur URLs im Datenset bleiben, die im Crawl und in der GSC bekannt sind. Das kann die Datensicht am Ende sehr verzerren.

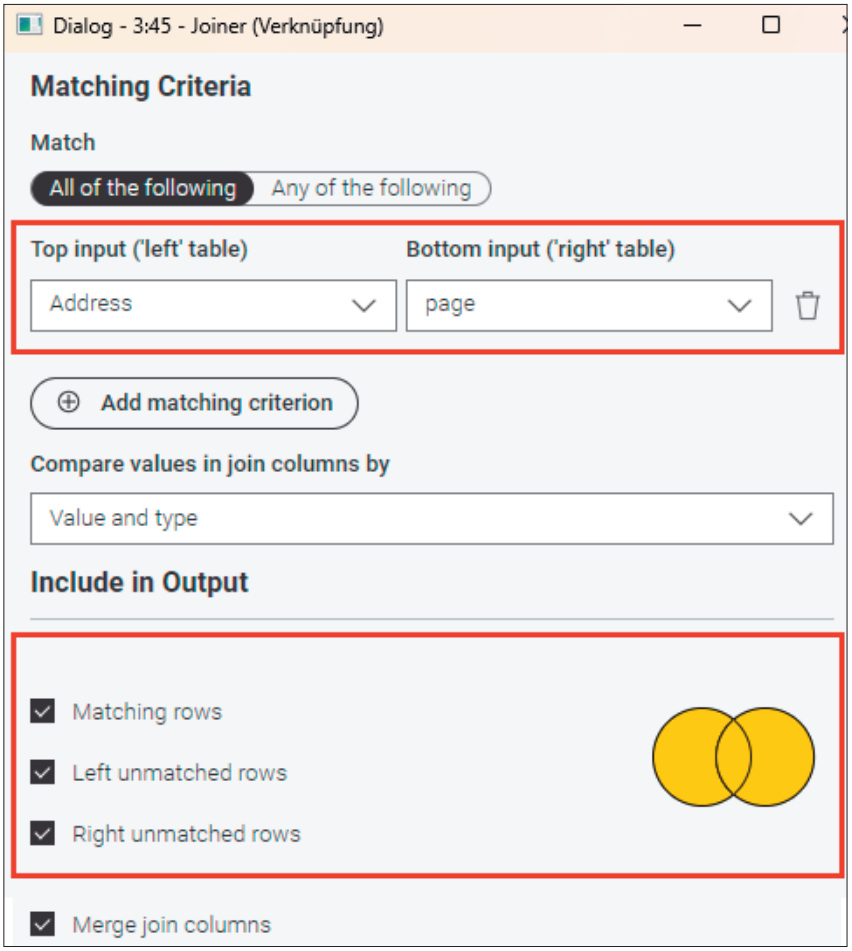


Abb. 6: Konfiguration des JOINERS

den sogenannten JOINER, das Äquivalent zum SVERWEIS() in Excel, zurückgegriffen.

Der JOINER hat im Gegensatz zu anderen Knoten zwei Eingangsports, um zwei Datensets miteinander zu kombinieren. Die Crawldaten werden mit dem oberen Port des JOINERS verbunden. Hierzu wird der letzte Knoten zur Aufbereitung der Crawl-Daten, der ROW FILTER, verwendet. Die GSC-Daten werden mit dem unteren Port verbunden (Abb. 5).

Damit die Verknüpfung der Daten funktioniert, wird wie im SVERWEIS() eine Referenzspalte benötigt. Dies ist hier die URL. In den Crawl-Daten nennt sich die passende Spalte „Address", in den GSC-Daten ist das die „page". Als Art des Matchings im JOINER wird der FULL OUTER JOIN gewählt, es werden also alle Daten miteinander verknüpft (Abb. 6).

Als Letztes wird noch die Einstellung „merge join columns" gewählt. Somit existieren im Datenset nicht zwei Spalten mit den gleichen Inhalten, den URLs. Außerdem ist das für den anschließenden Schritt der Segmentierung von Vorteil.

Erscheint die Ampel unter dem JOINER grün, kann die Segmentierung beginnen.

Die individuelle Segmentierung

Nun beginnt die Segmentierung der URLs. Dafür werden nun nach und nach alle URLs anhand ihrer Gemeinsamkeiten gruppiert.

Die erste Segmentierung soll die URLs verschiedenen Seitentemplates zuordnen. Somit sollen beispielsweise alle URLs, die laut ihrer Struktur ein Blogartikel sind, den Hinweis „Blogdetailseite" angehängt bekommen, während die Startseite dem Segment „Startseite" zugeordnet wird.

Das Ziel ist es, dass das Datenset eine neue Spalte bekommt, die für jede URL den Typ des Seitentemplates beinhaltet. Hierfür wird der Knoten RULE ENGINE benötigt. Dieser mächtige Knoten reichert ein Datenset mit weiteren Daten an.

Konfiguration der RULE ENGINE

Im Dialogfenster der RULE ENGINE befinden sich auf der linken Seite alle vorhandenen Spalten des Datensets (Abb. 7, Ziffer 1). Rechts daneben werden alle Funktionen angezeigt, die zur Anreicherung der Daten notwendig sind (Abb. 7, Ziffer 2). Entscheidend ist das

HINWEIS

In einer perfekten Welt mit einer perfekten URL-Struktur lässt sich das Datenset sehr leicht segmentieren. Die Praxis zeigt jedoch, dass viele URL-Strukturen historisch wachsen oder im Laufe der Zeit auf eine neue Sturktur gewechselt wird. Somit kann auch die Segmentierung an manchen Stellen kniffliger werden.

Feld Expression (Abb. 7, Ziffer 3). Hier werden alle Bedingungen eingegeben, damit die Daten entsprechend angereichert werden können. In grüner Schrift (durch // wurden die Zeilen auskommentiert, das heißt, sie werden bei der Ausführung nicht berücksichtigt. Dies kann man auch gut nutzen, um erklärende Kommentare zur Segmentierung einzufügen) wird hier bereits beispielhaft gezeigt, wie die Anforderungen daran zu schreiben sind:

» Je Zeile wird nur eine Bedingung definiert.
» Die Zeilen werden von oben nach unten verarbeitet. Greift demnach bereits eine Bedingung für einen Wert des Datensets, wird die nachfolgende Bedingung nicht mehr angewendet.
» Die letzte Zeile ist immer „TRUE => „default outcome“ und bedeutet, dass alle Daten, die nicht in die definierten Anforderungen fallen, mit dem Wert default outcome gekennzeichnet werden.

ACHTUNG: Keine Angst, falls die Angaben in der Theorie noch nicht ganz klar sind. Mit den nachfolgenden Beispielen wird das noch deutlicher.

Im unteren Teil des Knotens gibt es außerdem noch die Einstellung der Spalte. Hier wird angegeben, ob eine neue Spalte an das Datenset angehängt (= Append Column) werden soll und wenn ja, wie diese heißen soll. Alternativ kann auch eine bestehende Spalte überschrieben werden (= Replace Column) (Abb. 7, Ziffer 4).

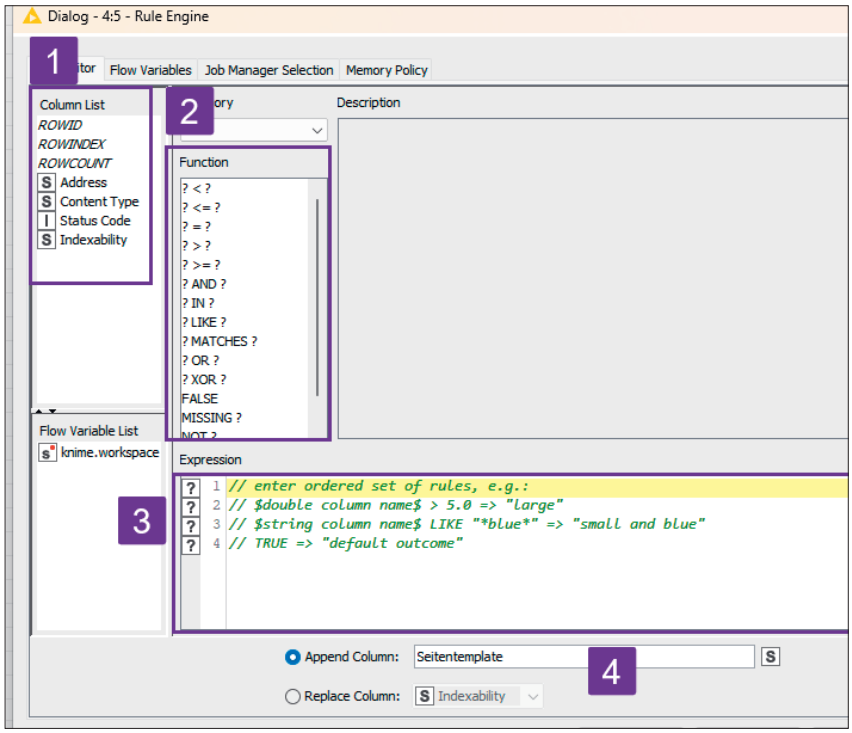


Abb. 7: Konfiguration der RULE ENGINE

Address=page	Seitentemplate
https://www.gettraction.de/	Startseite
https://www.gettraction.de/blog/kni...	Blogdetailseite
https://www.gettraction.de/blog/seo...	Blogdetailseite
https://www.gettraction.de/ueber-u...	Sonstige

Abb. 8: Ergebnis der Anforderungen aus der RULE ENGINE

Konkrete Einstellungen in der RULE ENGINE

Weil in diesem Anwendungsfall URLs anhand ihrer Struktur bestimmten Seitentemplates zugeordnet werden sollen, werden die Anforderungen anhand der URL-Spalte formuliert. Diese Spalte heißt nach der Verknüpfung im JOINER address=page.

Zum Start ist es sinnvoll, in die Vorschau des Datensets zu gehen, um erste Gemeinsamkeiten in der URL-Struktur ableiten zu können. Diese Vorschau wird über einen Rechtsklick auf den Knoten COLUMN FILTER und die Auswahl „Filtered table“ ersichtlich.

Erste Seitentemplates sind zum Beispiel Startseite und Blogdetailseite.

Die Expression kann wie folgt gestartet werden:

HINWEIS

Die Segmentierung von URLs ist ein iterativer Prozess. Denn im Rahmen der Bildung von Segmenten können immer wieder neue URLs auffallen, die noch keinem Muster zugeordnet sind. Zusätzlich kann zwischendurch immer wieder kontrolliert werden, ob die definierten Muster auch tatsächlich für alle gewünschten URLs greifen. Am Ende sollten so wenige Seiten wie möglich im Segment „Sonstige“ landen.

» $Address=page$ MATCHES „https://www.gettraction.de/?“=> „Startseite“
» $Address=page$ MATCHES „https://www.gettraction.de/blog/.*“ => „Blogdetailseite“
» TRUE => „Sonstige“

Stimmt die Anforderung eines Ein-

```
Expression
S 1 $Address=page$ MATCHES "https://www.gettraction.de/?"=> "Startseite"
S 2 $Address=page$ MATCHES "https://www.gettraction.de/(get-seo-intelligence|get-seo-intelligence|get-seo-success)/[^/]+/?" => "Leistungsseiten"
S 3 $Address=page$ MATCHES "https://www.gettraction.de/blog/?" => "Blogübersicht"
S 4 $Address=page$ MATCHES   "https://www.gettraction.de/blog/.*" => "Blogdetailseite"
S 5 $Address=page$ MATCHES "https://www.gettraction.de/ueber-uns/seo-jobs/.+" => "Jobdetailseiten"
S 6 $Address=page$ MATCHES "https://www.gettraction.de/ueber-uns/team/.+" => "Personendetailseiten"
S 7 $Address=page$ MATCHES "https://www.gettraction.de/(seo-darmstadt|seo-berlin)/" => "Standortseiten"
S 8 TRUE => "Sonstige"
```

Abb. 9: Erweiterung der RULE ENGINE um weitere Seitentemplates

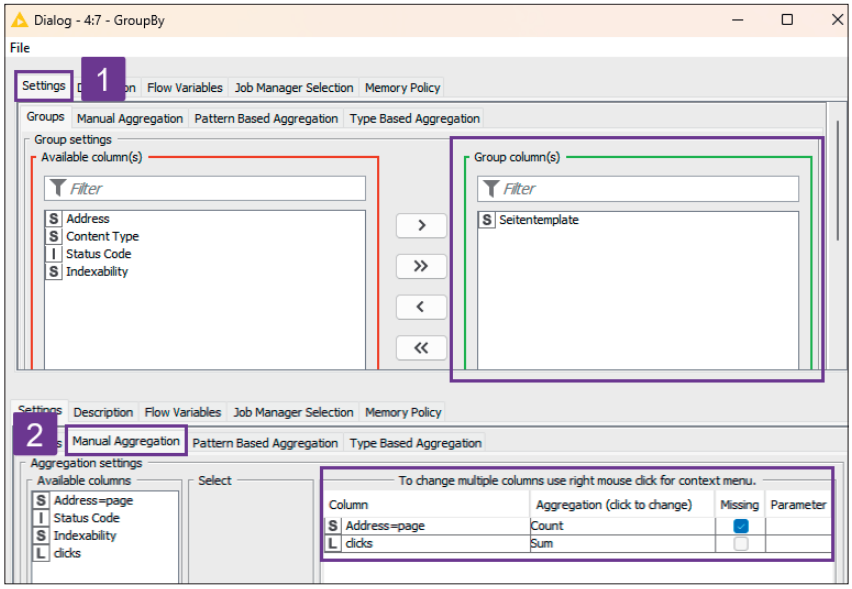
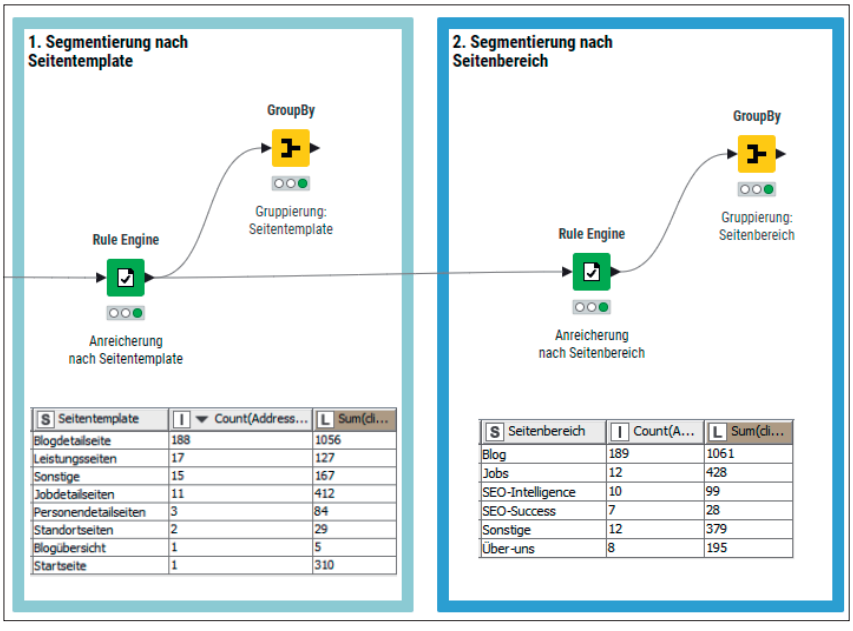


Abb. 10: Konfiguration im GROUP BY



Seitentemplate	Count(Address...	Sum(cli...
Blogdetailseite	188	1056
Leistungsseiten	17	127
Sonstige	15	167
Jobdetailseiten	11	412
Personendetailseiten	3	84
Standortseiten	2	29
Blogübersicht	1	5
Startseite	1	310

Seitenbereich	Count(A...	Sum(cli...
Blog	189	1061
Jobs	12	428
SEO-Intelligence	10	99
SEO-Success	7	28
Sonstige	12	379
Über-uns	8	195

Abb. 11: Vorschau der URL-Verteilung

Address=page	Seitentemplate	Seitenbereich
https://www.gettraction.de/get-seo-success/seo-fuer-verlage/	Leistungsseiten	SEO-Success
https://www.gettraction.de/get-seo-success/seo-audits-potentialanalysen/	Leistungsseiten	SEO-Success
https://www.gettraction.de/get-seo-success/content/	Leistungsseiten	SEO-Success
https://www.gettraction.de/get-seo-success/seo-maintenance/	Leistungsseiten	SEO-Success
https://www.gettraction.de/get-seo-success/seo-fuer-b2b-unternehmen/	Leistungsseiten	SEO-Success
https://www.gettraction.de/get-seo-success/seo-neukonzeption/	Leistungsseiten	SEO-Success
https://www.gettraction.de/get-seo-success/seo-im-relaunch/	Leistungsseiten	SEO-Success

Abb. 12: Vorschau des Datensets mit zwei Segmentierungen

trags in der Spalte mit einer der Anforderungen überein, wird in eine neue Spalte, die „Seitentemplate“ heißt, der Wert „Startseite“ bzw. „Blogdetailseite“ geschrieben. Stimmt der Eintrag nicht damit überein, wird als Wert in der neuen Spalte „Sonstige“ ergänzt.

Das erste Ergebnis dazu ist in Abb. 8 zu sehen. Es gibt nun eine neue Spalte, die als Einträge „Startseite“, „Blogdetailseite“ oder „Sonstige“ hat. Diese Vorschau ist durch einen Rechtsklick auf die RULE ENGINE und die Auswahl „Classified values“ möglich.

Nun kann die Segmentierung weitergehen. Ziel ist es, am Ende so wenige URLs wie möglich mit dem Wert „Sonstige“ zu erhalten. Deshalb wird nun Schritt für Schritt, je nach identifiziertem Muster, ein Seitentemplate definiert.

Kleiner Trick zum Segmentieren:
Um die URL-Muster genau zu definieren, helfen Kenntnisse in RegEx in jedem Fall. Im Jahr 2025 reicht es aber auch, auf ChatGPT oder eine andere schlaue KI zurückzugreifen, um einfache reguläre Ausdrücke zu beschreiben und einfach in die Expression der RULE ENGINE zu kopieren.

Nachdem die größten Muster gefunden worden sind, sieht das Ganze in der Expression so aus (Abb. 9).

Zur besseren Übersicht wird an die RULE ENGINE ein weiterer Knoten angehängt, der GROUP BY. Dieser Knoten ist das Äquivalent zur Pivot-Tabelle und kann Daten zur besseren Übersicht gruppieren.

Um besser sehen zu können, wie viele URLs im jeweiligen Seitentemplate getroffen worden sind, werden folgende Einstellungen getätigt:

» In das rechte Fenster Group columns wird die Spalte „Seitentemplate“ gezogen (Abb. 10, Ziffer 1).
» Im Reiter Manual Aggregation wird die Spalte Address und als Aggregation Count ausgewählt. Die Spalte clicks erhält die Aggregation Sum (Abb. 10, Ziffer 2).

Mit der Vorschau im GROUP BY wird sichtbar, wie viele URLs zu den definierten Seitentemplates gehören und wie hoch die Klicks in den Bereichen sind (Abb. 11).

Ist die Segmentierung nach dem Seitentemplate abgeschlossen, ist es auch möglich, die Daten über einen weiteren Knoten erneut einer Segmentierung zu unterziehen, beispielsweise nach Seitenbereichen. Dazu wird einfach eine weitere RULE ENGINE an den bestehenden Workflow angehängt, um noch einmal auf eine andere Art und Weise zu segmentieren. Diese Segmentierung ist sehr individuell. Es gibt dabei auch die Möglichkeit, URLs nach B2C- und B2B-Zielgruppe aufzuteilen, was anhand der URL-Struktur nicht zwingend ableitbar ist.

Hier ist je nach Analyseansatz eine Vielzahl an Segmentierungen möglich.

Beispiele für weitere Anwendungsfälle der Segmentierung

» Für einen Shop soll herausgefunden werden, in welchen Produktsegmenten die meisten Klicks erzielt werden, gleichzeitig soll aber auch abgeleitet werden, ob es sich dabei um Produktdetailseiten oder Kategorieseiten handelt.
» Für eine Unternehmensseite sollen Argumente gefunden werden, einen Pressebereich abzuschalten, in dem die meisten URLs liegen, aber die wenigsten Klicks erzielt werden.
» Es kann eine Vielzahl von Flag-Spalten gebildet werden, die als Wert nur „Ja“ oder „Nein“ aufweisen, zum Beispiel, ob die URL eine Pagination, Parameter beziehungsweise Umlaute aufweist oder ob es sich um eine AMP-URL handelt.

Die Anzahl der möglichen Segmente ist somit fast unendlich. Alles, was einem in der späteren Analyse dabei hilft, URL-Sets getrennt voneinander/miteinander zu vergleichen, kann als Segment angelegt werden.

Fazit

Obwohl der Workflow zunächst sehr einfach aussieht, können damit doch mächtige Analysen gestartet werden. Denn häufig ist diese Segmentierung erst der Anfang, um dann immer tiefer in die Seitenstruktur vorzudringen.

Vielleicht ist diese Segmentierung ja ein guter Start für eine Aufräumaktion, denn ein SEO-Frühjahrsputz kann so mancher Website neue Power verleihen. Viel Spaß beim Ausprobieren.

PS: Der fertige Workflow ist wie immer im KNIME-Hub unter einfach.st/knime86 verfügbar und ist per Drag-and-drop in die eigene KNIME-Oberfläche einlesbar. Bei Fragen und Anmerkungen zum Workflow kann gerne ein Austausch via E-Mail oder LinkedIn stattfinden.

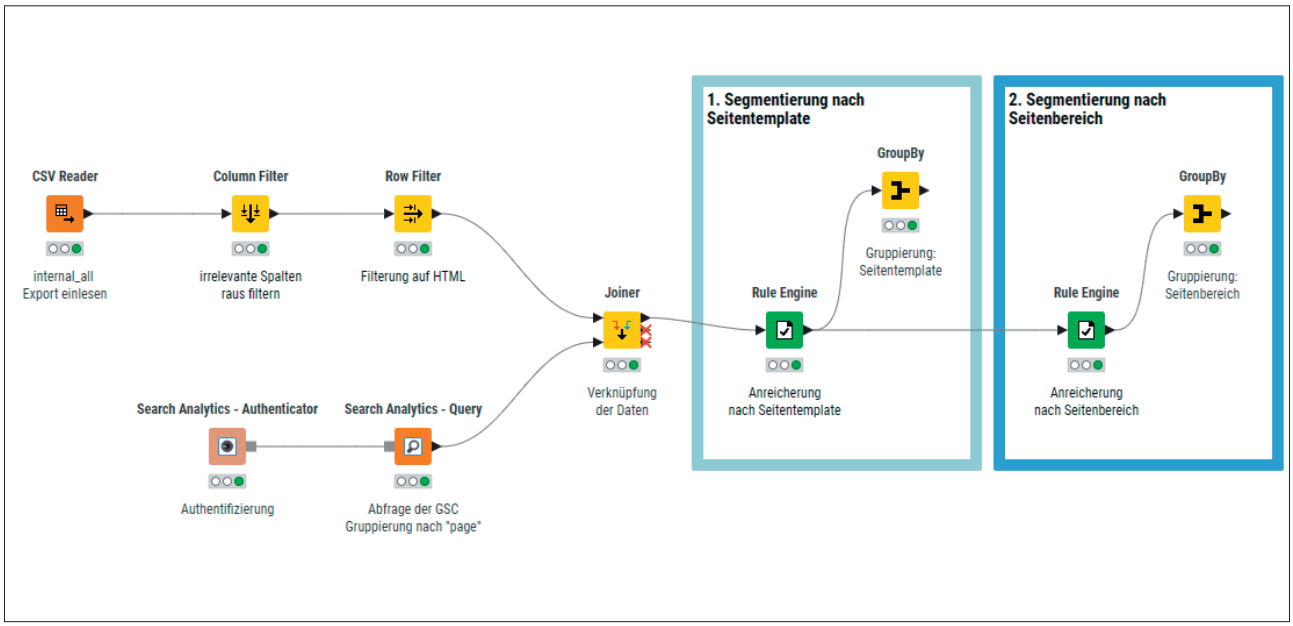


Abb. 13: Gesamter KNIME-Workflow