Rebecca Schwarz

Aufräumaktion mit KNIME: Wie viele URLs sind für 80 % der Klicks verantwortlich?

In der letzten Ausgabe lag der Fokus auf der Segmentierung einer Website nach Seitenbereichen. Das Ziel war es, sich eine Übersicht zu verschaffen, wie viele URLs in den einzelnen Seitenbereichen liegen und wie viele Klicks und Impressionen sich je Bereich ergeben. Nun geht Rebecca Schwarz eine Ebene tiefer in die Segmentierung und beschäftigt sich mit der Frage: Welche URLs sind für den größten Anteil des Traffics im Seitenbereich verantwortlich und welche URL generieren kaum bis keinen Traffic. Mit anderen Worten: Welche URL sind die berühmten „heiligen Kühe", bei denen man bei allen Änderungen ganz besonders umsichtig sein sollte.
Das Ergebnis der Analyse ist eine Löschliste, um die Website von irrelevanten URLs zu befreien oder zumindest, um mit den Verantwortlichen zu diskutieren, welche unnötige Last von Dokumenten auf dem Server liegt. Und bekanntlich berichten viele Unternehmen von gestiegenen Rankings, nachdem man unnötige, weil von Besuchern verschmähte Seiten gelöscht wurden. Insofern kann so eine Liste ein wertvolles Instrument für die Suchmaschinenoptimierung darstellen.

Auf geht's in den Frühjahrsputz 2025!

Ideal vs. Realität von Website-Strukturen

Die ideale Struktur einer Website sind schön strukturierte Verzeichnisse, die sinnvoll benannt sind, und Detailseiten, die einem interpretierbaren Muster folgen. Die Auswertbarkeit ist kinderleicht und alle URLs sind anhand ihrer Struktur einwandfrei den Verzeichnissen zuzuordnen. Was für eine schöne Vorstellung! Doch die Realität zeigt meist das Gegenteil: Historisch gewachsene URL-Strukturen, keine stringente Benennung und ganze Verzeichnisse, die verstauben und den Verantwortlichen kaum noch bekannt sind.

Um sich eine erste Übersicht zu verschaffen, ging es in Ausgabe #91 deshalb um die Segmentierung einer Website nach Seitenbereichen und Seitentypen. So konnte abgeleitet werden:

» welche Seiten existieren,
» wie viele URLs je Seitenbereich vorhanden sind und
» wie sich die Klicks und Impressionen laut Google Search Console (kurz: GSC) auf die Segmente verteilen.

Nun wird noch einen Schritt weiter in die Daten gedrillt. Es geht dabei um die Fragen: Welche der URLs sind hauptsächlich für den Traffic verantwortlich. Denn oft gibt es einzelne sehr klickstarke Seiten und wenn sie einem gemeinsamen Bereich zugeordnet sind, erscheint dieser sehr erfolgreich - obwohl der Großteil der Seiten in dem Bereich kaum bis keinen Traffic erzeugen. Aber eins nach dem anderem!

KNIME als mächtiges Auswertungstool

Auch für diese Analyse ist die Open-Source-Software KNIME das Tool der Wahl. Das liegt nicht nur an der kostenfreien Nutzungsmöglichkeit und einer großen Community, die immer wieder neue Anwendungsfälle zur Verfügung

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 im redaktionellen SEO. Um größere Datenmengen effizient zu verarbeiten und bei wiederkehrenden SEO-Tasks Zeit zu sparen, nutzt sie die Open-Source-Software KNIME.

stellt, sondern auch an der sehr leichten Bedienbarkeit.

Denn KNIME arbeitet mit sogenannten Knoten, mit denen Daten verändert, angereichert oder zusammengefasst werden können. Im Laufe einer Analyse werden verschiedene Knoten nach und nach miteinander verbunden. Durch den damit entstandenen Workflow können dann Daten hindurchfließen. Der große Vorteil ist, dass jederzeit einfach nachvollzogen werden kann, was je Knoten mit den Daten passiert. Außerdem sind einmal erstellte Workflow immer mit gleichförmigen Daten nutzbar.

Die Einarbeitung ins Tool lohnt sich in jedem Fall, denn die Anwendungsmöglichkeiten sind unendlich. Übrigens sind alle vorgestellten Workflows der Website Boosting online abrufbar und für die eigene KNIME-Umgebung nutzbar. Alle vorgestellten Workflows finden Sie über den Link ganz am Ende des Beitrags.

Start der Analyse und Wiederholung der groben Segmentierung

Um mit der Analyse zu starten, wird zunächst wieder Folgendes benötigt:

» eine Liste aller URLs einer Domain (am besten: internal_all Export aus dem ScreamingFrog)
» die Installation der KNIME-Umgebung
» Zugang zur Google Search Console der Website

Die Vorbereitung und Anreichung der Daten ist zunächst einmal der gleiche Workflow wie in der vergangenen Ausgabe. Deshalb erfolgt die Erklärung dazu einmal im Schnelldurchlauf.

Schritt eins: Einlesen der Daten

Zunächst werden die Crawl-Daten ins Datenanalyse-Tool KNIME eingelesen. Das funktioniert mit dem Knoten CSV-READER. (Abbildung 1,1) Dazu kann der internal_all Export aus dem ScreamingFrog einfach per Drag and Drop in die Tool-Oberfläche gezogen werden. Der passende READER-Knoten erscheint und es öffnet sich ein Dialogfenster.

In der Konfiguration wird im Reiter „Encoding“ UFT-8 oder UTF-16 gewählt. Mit Klick auf „OK“ schließt sich der Knoten. Nun wird mit Rechtsklick auf den Knoten „Execute“ gewählt, um den Knoten auszuführen. Dies ist bei jedem anderen Knoten auch der Weg, um die Daten nach der gewählten Konfiguration des Knotens zu bearbeiten.

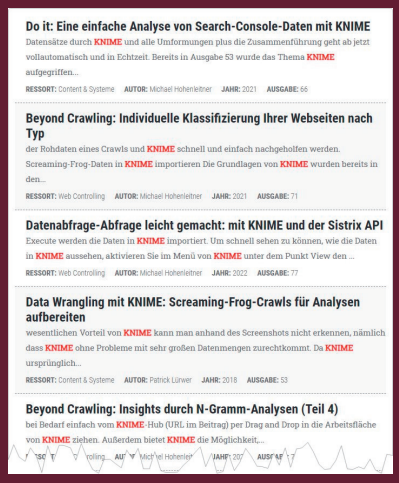
Schritt zwei: Entfernung irrelevanter Spalten

Nun werden mit dem nächsten Knoten, dem COLUMN-FILTER (Abbildung 1,2), alle Spalten aus dem Datenset entfernt, die für die Analyse nicht notwendig sind. Wird der Knoten mit dem CSV-READER verbunden, stehen zunächst alle Spaltennamen in der Konfiguration im rechten Fenster „Includes“. Über einen Doppelklick können alle irrelevanten Spalten entfernt werden.

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.



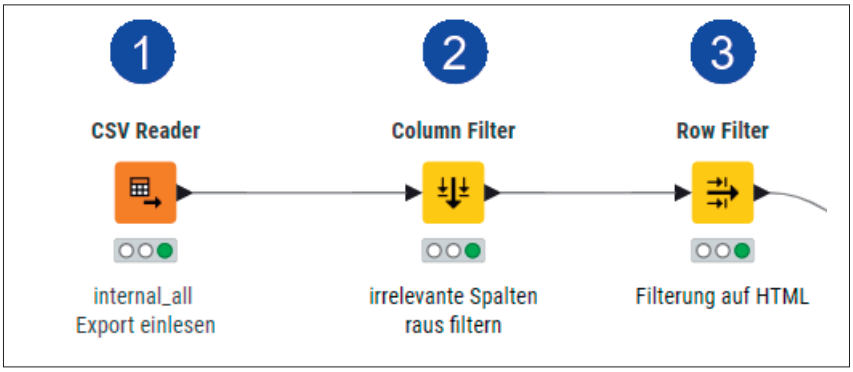


Abb. 1: Vorbereitung der Daten

Benötigt werden für die Analyse nur die Spalten Address, Status Code, Content Type und Indexability. Auch hier muss der Knoten über „Execute“ einmal ausgeführt werden, damit die Bearbeitung im nächsten Knoten weitergehen kann.

Schritt drei: Filterung auf HTML-Dokumente

Für die Betrachtung der URLs interessieren in dieser Analyse nur die HTML-Dokumente. Deshalb wird das Datenset mithilfe des Knotens ROW FILTER auf diesen Content-Typus gefiltert (Abbildung 1,3).

In der Konfiguration wird als Filter column „Content Type“ und als Operator „Matches Wildcard“ gewählt, als Value wird „*html*“ eingegeben. Wichtig ist zusätzlich, dass als Case matching „Case insensitive“ gewählt wird, damit wirklich alle Dokumente getroffen werden, die als HTML gekennzeichnet sind.

Nachdem die Daten nun vorbereitet sind, wird das Datenset mit den Daten der Google Search Console angereichert.

Schritt vier: Anreicherung mit den GSC-Daten

Dazu gibt es mittlerweile einen neu entwickelten Knoten von Mario Fischer, der die Daten direkt aus der GSC abholt und die KNIME-Umgebung importiert. Ganz ohne lästigen händischen Export und Import.

Dazu wird zum einen der Knoten SEARCH ANALYTICS – AUTHENTICATOR (Abbildung 2,1) benötigt. Dieser wird einfach in die Umgebung gezogen und zunächst nicht mit dem aktuellen Workflow verbunden. Denn erst werden die Daten abgefragt. In der Konfiguration kann im Drop-Down-Menü die Authentifikationsdauer gewählt werden. Mit der Ausführung des Knotens wird nun direkt das aktive Browser-Fenster geöffnet. Vor der Ausführung sollte deshalb der Browser mit dem verknüpften GSC-

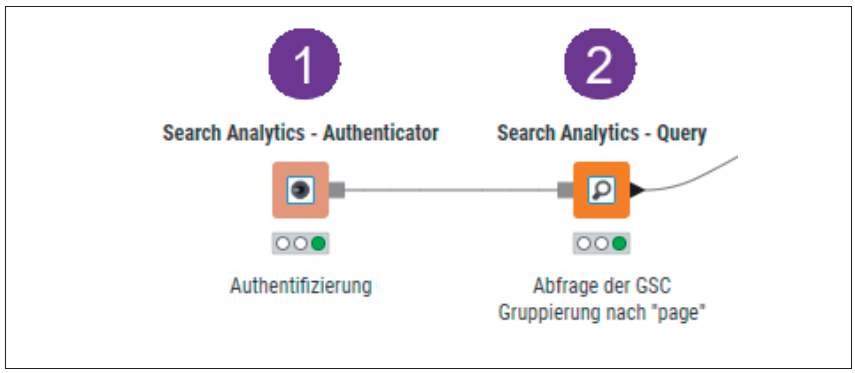


Abb. 2: Integration der GSC-Daten in KNIME

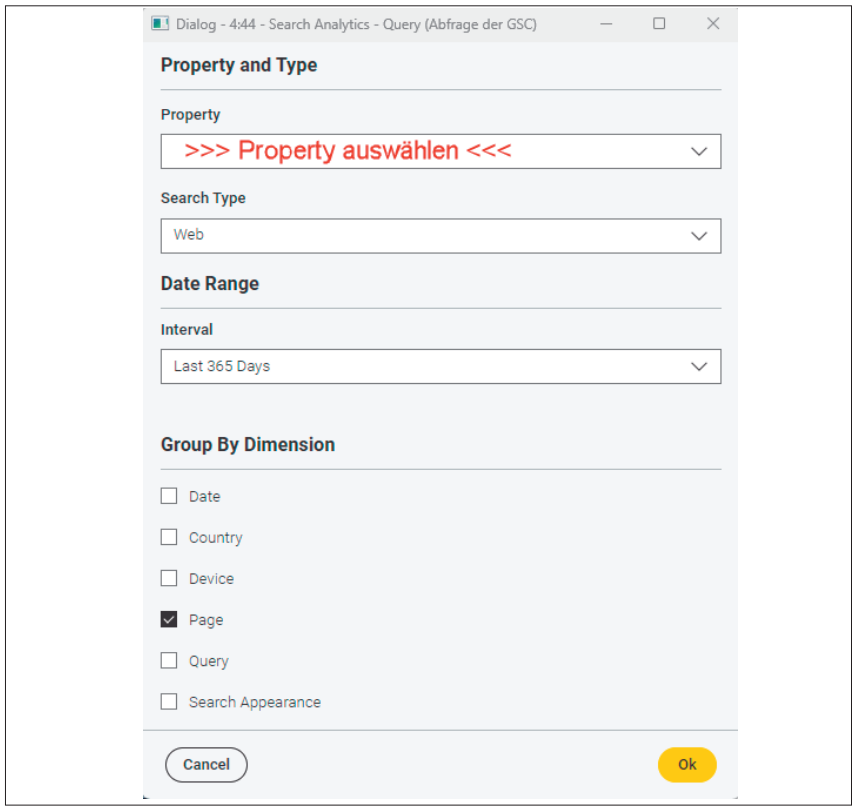


Abb. 3: Konfiguration zur Abfrage der GSC-Daten in SEARCH ANALYTICS – QUERY

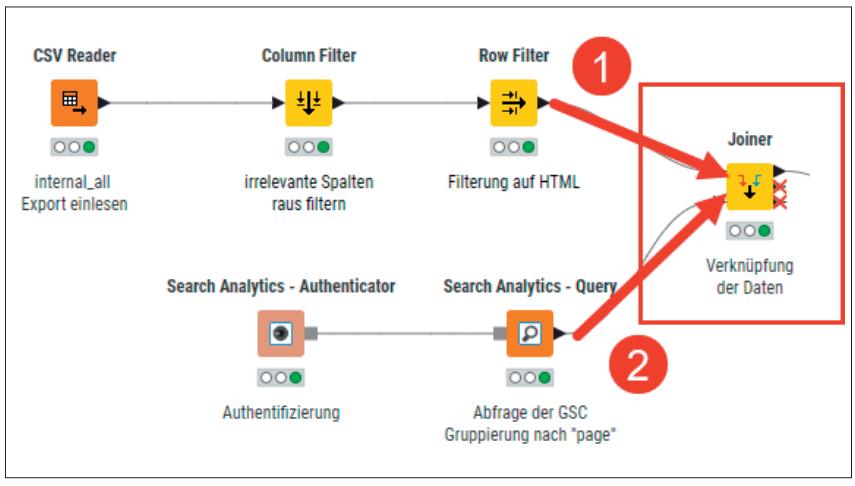


Abb. 4: Verknüpfung der Daten über den JOINER

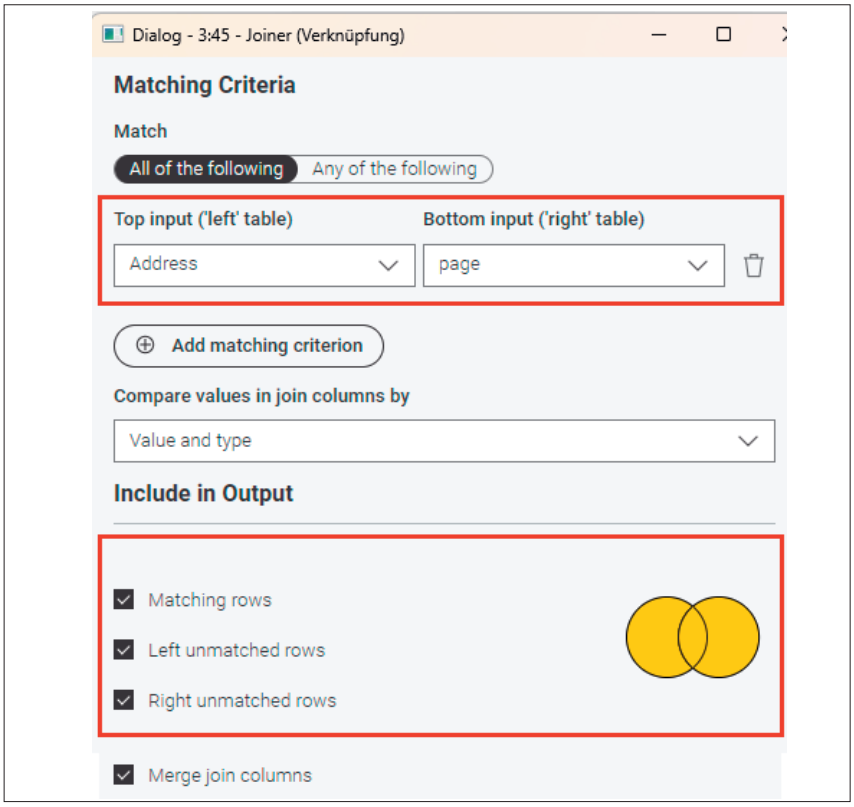


Abb. 5: Konfiguration im JOINER

Zugang geöffnet sein. Ansonsten kann die URL aus dem geöffneten Browser auch entsprechend in einen anderen Browser kopiert werden. Ist das passende Profil ausgewählt, ist die Konfiguration abgeschlossen.

Verbunden wird nun der Knoten SEARCH ANALYTICS – QUERY (Abbildung 2,1). Hier wird nun ausgewählt, welche Daten aus der GSC geholt werden sollen. Die aktuelle Konfiguration ist in Abbildung 3 zu sehen.

HINWEIS ZU DEN DATEN

Der FULL OUTER JOIN hat zur Folge, dass es drei verschiedene Datentöpfe gibt:

1. URLs, die im Crawl vorhanden, aber der Google Search Console (GSC) unbekannt sind.
2. URLs, die in der GSC bekannt sind, aber nicht mehr im Crawl, also in der internen Verlinkung, existieren.
3. URLs, die in der GSC und in den Crawldaten bekannt sind.

Je nach Datenlage kann es daher sein, dass es mitunter keine Informationen in den entsprechenden Spalten gibt. Denn URLs, die der GSC nicht bekannt sind, haben selbstverständlich keine Klickdaten. Während URLs, die nicht in den Crawldaten enthalten waren, keine Informationen zum Indexierungsstatus oder dem Statuscode haben.

TIPP

Auf Kriegsfuß mit RegEx-Ausdrücken?
Kein Problem! Nutzen Sie einfach ChatGPT oder Gemini und beschreiben Sie dort, was Sie genau filtern möchten, und Sie erhalten einen komplett fertigen RegEx-Ausdruck zurück, den Sie einfach nur in das Feld der entsprechenden Filternode einkopieren.
Hier ein Beispiel für so einen Prompt: „Mach mir eine RegEx, die mir alle URL filtert, die mehr als drei Verzeichnisebenen haben und die nicht mit der Dateiendung .pdf enden." Das Ergebnis sieht recht kompliziert aus, aber es funktioniert: ^(?:https?:\/\/)?[^\/]+(?:\/[^\/]+){4,}(?![^\/]*\.pdf)$

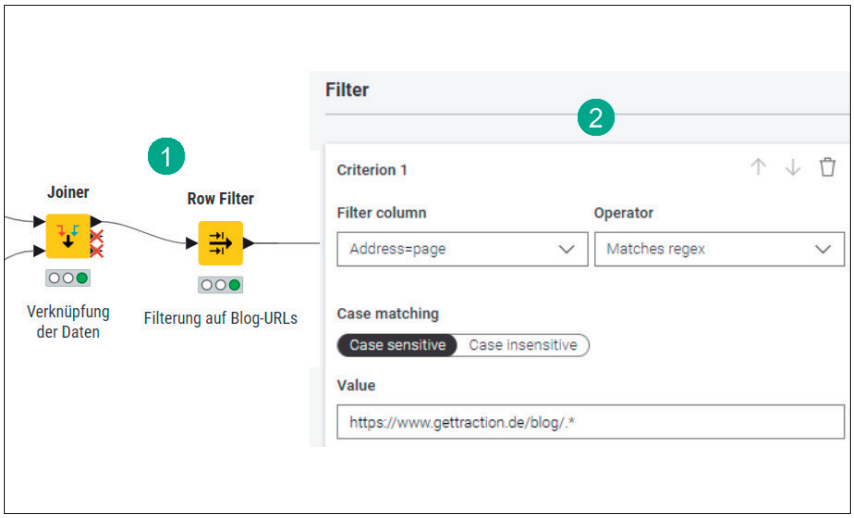


Abb. 6: ROW-FILTER hinzufügen und konfigurieren

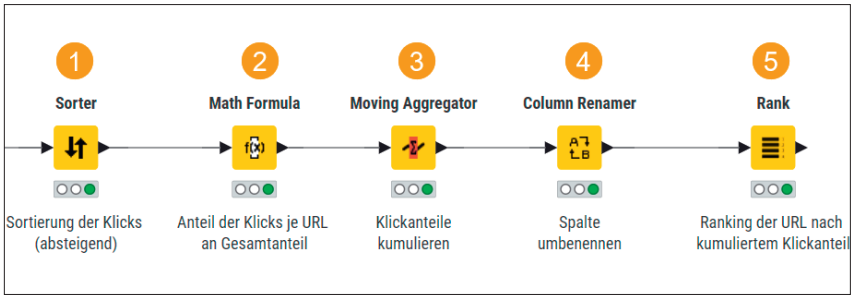


Abb. 7: Analyse der Klickverteilung

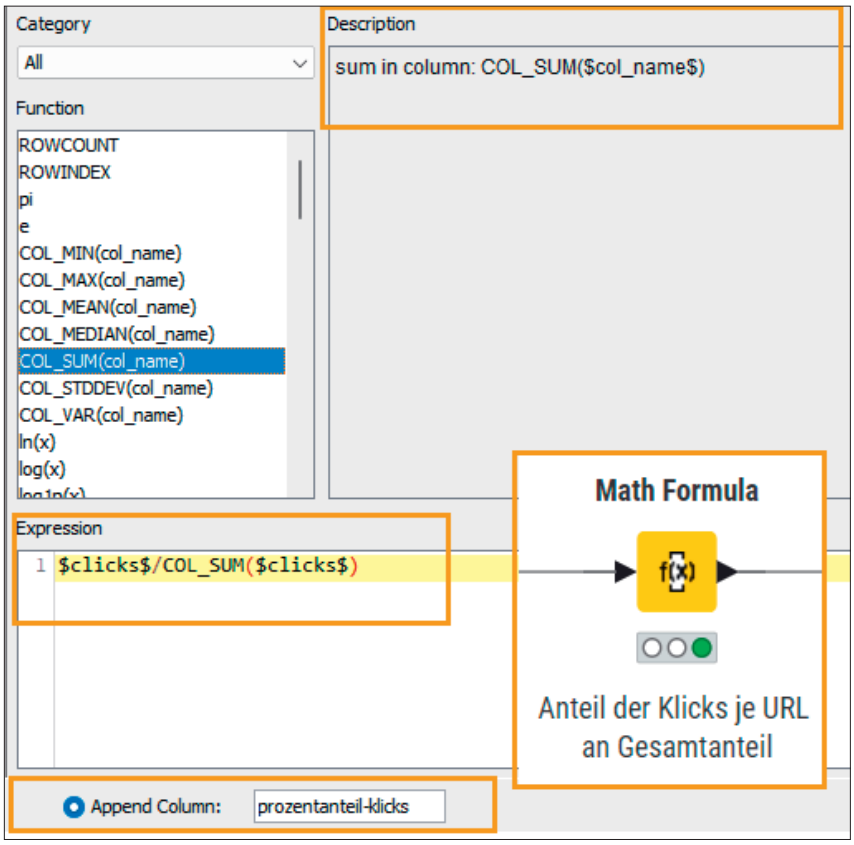


Abb. 8: Konfiguration in MATH FORMULA

Schritt fünf: Verknüpfung der Daten

Um die GSC-Daten mit den aufbereiteten Crawldaten zu kombinieren, kommt jetzt der sehr mächtige Knoten JOINER zum Einsatz. Es ist das Äquivalent zum SVERWEIS() in Excel. In den JOINER werden dazu beide Datenstränge zusammengefügt. Zum oberen Port des Knotens werden die vorbereiteten Crawldaten verbunden. (Abbildung 4,1) Am unteren Port werden die GSC-Daten angebunden. (Abbildung 4,2)

Für die Verknüpfung wird in der Konfiguration Folgendes ausgewählt (Abbildung 5):

» Wahl der Referenzspalte in Top input „Address" und Bottom input „page".
» Als Matching-Methode werden der FULL OUTER JOIN und somit alle Kombinationen der Daten gewählt.
» Zusätzlich wird die Checkbox „merge join columns" angeklickt. Diese Einstellung fasst die Referenzspalte der beiden Datensets zusammen. Deshalb wird aus der Spalte „page" und der Spalte „Address" die neue Spalte „Address = page".

Bei erfolgreicher Ausführung sind nun beide Datensets vereinigt und alle URLs, die auch der GSC bekannt sind, haben Daten zu Klicks und Impressionen erhalten.

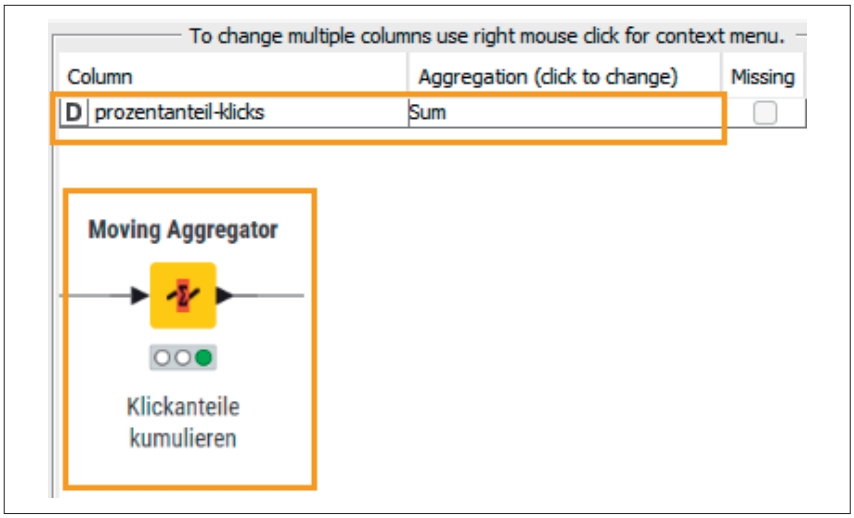


Abb. 9: Konfiguration von MOVING AGGREGATOR

Address=page	kumulierter-anteil-klicks	ranking
https://www.gettraction.de/blog/knim	24%	1
https://www.gettraction.de/blog/seo-	45%	2
https://www.gettraction.de/blog/pres	51%	3
https://www.gettraction.de/blog/goog	56%	4
https://www.gettraction.de/blog/goog	60%	5
https://www.gettraction.de/blog/richt	65%	6
https://www.gettraction.de/blog/goog	69%	7
https://www.gettraction.de/blog/reca	73%	8
https://www.gettraction.de/blog/das-	76%	9
https://www.gettraction.de/blog/reca	78%	10
https://www.gettraction.de/blog/goog	80%	11

Abb. 10: Liste der URLs mit Ranking

Start in die tiefere Analyse der URLs

Nun kann mit der Segmentierung der URLs begonnen werden. In Ausgabe #91 wurde gezeigt, welche Gemeinsamkeiten URLs haben, um diese anhand ihrer Struktur zu clustern. Somit wurden dann alle URLs, die im Verzeichnis /blog/ liegen, dem Segment „Blogseiten“ zugeordnet, während Seiten mit dem Verzeichnis /team/ als „Personenseiten“ klassifiziert wurden.

Schlussendlich kam dabei eine Ansicht heraus, die zeigt, wie viele URLs je Segment vorhanden sind und wie viele Klicks und Impressionen in den Segmenten hängen. Im Zusammenhang konnte erkannt werden, dass im Blogverzeichnis die meisten Klicks generiert werden, jedoch auch die meisten URLs bestehen. Deshalb kann sich nun tiefer in dieses eine Verzeichnis gedrillt werden.

WICHTIGER HINWEIS

In der Vorschau des Datensets können die Daten jeder Spalte auch jederzeit ab- oder aufsteigend sortiert werden. Diese Sortierung findet jedoch nur in der Vorschau statt und hat keine Auswirkung auf das Datenset.

Mögliche Anwendungsbeispiele:

» Als Publisher möchte man sich einen Überblick verschaffen, welche Artikeldetailseiten für den hauptsächlichen Traffic verantwortlich sind.

» Für den Ratgeberbereich einer Website soll eruiert werden, welche Einträge optimiert oder neu verfasst werden müssen.

Für die tiefere Analyse wird zunächst das formale Muster definiert, anhand dessen alle URLs gematcht werden können. Wenn bereits eine Segmentierung nach Seitenbereich stattgefunden hat, kann man natürlich diese Muster einfach übernehmen. Ansonsten kann mit einem cleveren Prompt an die KI oder den eigenen RegEx-Kenntnissen schnell ein passendes RegEx Muster erzeugen und damit filtern.

Für Blogdetailseiten wie: *https://www.gettraction.de/blog/seo-plugins-fuer-chrome/* ist das passende Muster mit RegEx: *https://www.gettraction.de/blog/.**. Um die URL auf diese Muster hinzufiltern, wird an den JOINER ein einfacher ROW-FILTER angehängt (Abbildung 6,1). In der Konfiguration (Abbildung 6,2) wird als Filter column „Address = page“ und als Operator „Matches regex“ eingetragen.

Als Value wird *https://www.gettraction.de/blog/.*“* eingegeben. Nach der Ausführung befinden sich im Datenset nun nur noch Blogdetailseiten.

Wie viele URLs sind für den Hauptteil der Klicks verantwortlich?

Nun besteht schon eine bessere Übersicht. Was im hier gezeigten Beispiel sehr gering wirkt, ist in der Realität häufig deutlich massiver. Denn gerade bei Publishern mit Tausenden Seiten, kann die Analyse schnell auch mal 10.000 bis 20.000 Artikeldetailseiten betreffen. Um hier die Übersicht nicht zu verlieren, soll nun analysiert werden, wie viele URLs ausreichen, um den Hauptteil des Traffics zu generieren.

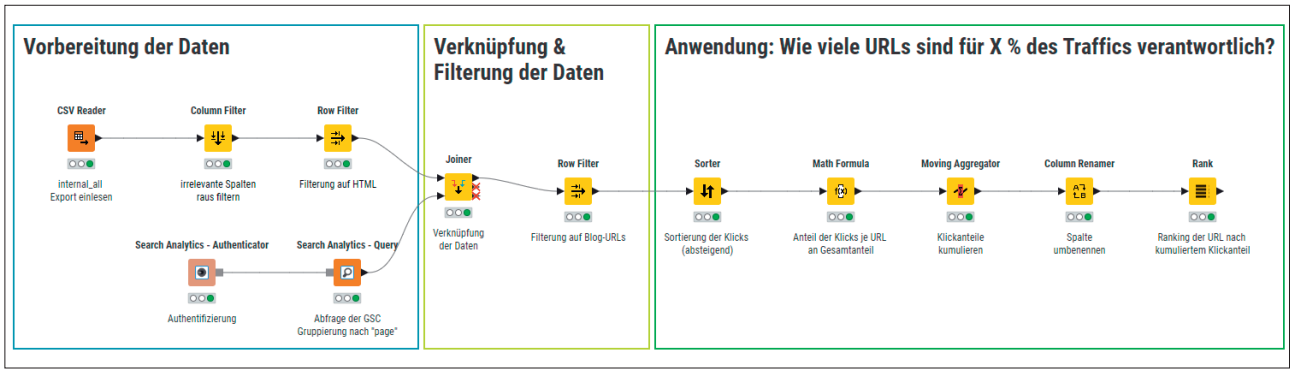


Abb. 11: Gesamter KNIME-Workflow zum Download im KNIME-Hub

Dazu werden fünf weitere Knoten benötigt, die nacheinander an den ROW-FILTER angehängt werden (Abbildung 7).

Schritt eins: SORTER

Mithilfe des Knotens SORTER werden die URLs absteigend nach Klicks sortiert. Dazu wird in der Konfiguration einfach als column clicks gewählt und dann als Order „descending“ geklickt (Abbildung 7,1)

Schritt zwei: MATH FORMULA

Mit diesem Knoten kann mit einer mathematischen Formel eine Berechnung durchführen. MATH FORMULA wird hier verwendet, um den Anteil der Klicks jeder einzelnen URL am Gesamtanteil der Klicks zu berechnen. In der Konfiguration sind einige Formeln vorgeschlagen (Abbildung 8).

Für die Anteilsberechnung wird $clicks$/COL_SUM($clicks$) in das Expression-Feld eingetragen. Denn mit der Formel COL_SUM($clicks$) wird zunächst die Gesamtzahl aller Klicks ausgerechnet (COL_SUM = Summe aller Spaltenwerte), durch die dann die Anzahl der Klicks der jeweiligen URL geteilt werden kann, um ihren prozentualen Anteil an den Gesamtklicks zu bestimmen und in eine neue Spalte mit dem Namen „prozentanteil-klicks“ zu schreiben.

Schritt drei: MOVING AGGREGATOR

Mit diesem Knoten werden die Anteilwerte nun kumuliert. In der Konfiguration wird die Spalte „prozentanteil-klicks“ hinzugefügt und die Aggregation „Sum“ gewählt (Abbildung 9). Nach erfolgreicher Ausführung ergibt sich eine weitere Spalte, die die prozentualen Anteile Zeile für Zeile aufsummiert.

Somit kann sehr einfach abgelesen werden, wann ein bestimmter Prozentwert erreicht wird. Dies wird mit den nächsten beiden Knoten noch deutlicher.

Schritt vier: COLUMN RENAMER

Hier geht es nur darum, die neu hinzugekommene Spalte neu zu benennen., da der Knoten eine eher kryptische Bennenung ausgibt. Aus dem Spaltennamen SUM(prozentanteil-klicks) wird somit kumulierter-anteil-klicks (Abbildung 7,4)

Schritt fünf: RANK

Mit dem Knoten RANK werden die URLs nun noch aufsteigend nach ihrem prozentualen Anteil an den Gesamtklicks gerankt. Die URL mit dem größten prozentualen Anteil erhält den Rang 1, die URL mit dem zweitgrößten Anteil den Rang 2 etc. (Abbildung 7,5).

Die daraus entstandene Liste kann nun einfach mit dem Knoten EXCEL WRITER als Excel-Datei exportiert werden. Mit einer kleine Anpassung der Kommazahlen als Prozentwerte lässt sich sehr leicht ablesen, dass beispielsweise elf URLs genügen, um auf 80 % der Klicks zu kommen (Abbildung 10). Bei weitaus größeren Websites ergeben sich hier häufig große Überraschungen. So wird schnell deutlich, dass bereits 200 URLs für 75 % des Traffics verantwortlich sind, obwohl über 5.000 URLs existieren.

Am Ende dieser Analyse geht es dann um die Entscheidung: Was wird aus all den URLs, die kaum bis gar nichts zum Traffic beitragen? Optimieren, löschen, einfach behalten?

Hinweis: Für diese Entscheidung müssen häufig weitere Faktoren berücksichtigt werden. So sollte vor der finalen Entscheidung geprüft werden, ob die URLs über andere Kanäle Traffic bringen oder ob es andere wertvolle Informationen je URL gibt, wie die Anzahl der Neuabonnenten oder die Anzahl der Backlinks. Um den Stein für eine Löschaktion ins Rollen zu bringen und kritischer zu prüfen, ist diese Ansicht aber eine sehr simple Möglichkeit.

Fazit

Einmal aufgesetzt ist dieser Workflow sehr leicht für jede Art von Website durchführbar. Es können schnell Erkenntnisse über die Website erlangt werden und ein spannender Startpunkt für die nächste große Aufräumaktion sein. Viel Spaß beim Ausprobieren!

Der fertige Workflow ist wie immer im KNIME-Hub unter *einfach.st/knime86* verfügbar und 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. ¶