Arne-Christian Sigge

A fool with a tool is still a fool?
Die größten Fehler bei der Anwendung von WDF*IDF-Tools

Die sechs Buchstaben „WDF-IDF“ haben letztes Jahr nicht nur in Diskussionen für Aufruhr in der Branche gesorgt. Die teils eilig aus dem Boden gestampften Tools erfüllten dem Vernehmen nach nicht immer die Erwartungen der Experten. Der steigenden Beliebtheit dieser Tools tat und tut dies jedoch keinen Abbruch. Umgekehrt ist einer der größten Fehler, die unerfahrene „Onliner“ oder SEOs bei der Erstellung von Text-Content machen können, die unkritische und unbedarfte Nutzung solcher WDF*IDF-Tools. Erforderlich sind gutes Hintergrundwissen und auch die Bereitschaft, sich mit den gelieferten Daten intensiv auseinanderzusetzen. Wer die Tools lediglich dazu nutzt, Texte zwischen zwei Kurven zu prügeln, hat grundsätzliche Dinge nicht verstanden und verschenkt dadurch wertvolle Potenziale beim Ranking. Arne-Christian Sigge klärt in deutlichen Worten die populärsten Irrtümer auf und zeigt praxisorientiert, in welchem gedanklichen Rahmen die sechs Buchstaben verwendet werden sollten.

Viele unerfahrene SEOs oder Kunden von SEO-Agenturen wenden WDF-IDF-Tools an wie ein Lehrerehepaar, das mit ergoogeltem Halbwissen und einer Wasserwaage die Handwerker im entstehenden Eigenheim in den Wahnsinn treibt, und vergessen dabei, dass jedes Werkzeug nur so gut ist wie der Handwerker, der es benutzt. So erhalten auch immer wieder Autoren abstruse Textaufträge von überforderten Auftraggebern, die glauben, ein WDF*IDF-Analyse-Tool ermögliche per Knopfdruck den Durchmarsch auf die ersten Plätze in den Suchergebnislisten bei Google & Co.

Ein Rückblick auf die Anfänge von WDF*IDF-Analysen und die Ursachen vieler Probleme in der Anwendung

Um zu verstehen, warum die Ergebnisse von WDF*IDF-Analysen mit Vorsicht zu genießen sind, lohnt ein Blick in die Vergangenheit, auf das ursprüngliche Anwendungsgebiet der WDF*IDF-Technologie.

Die ersten WDF*IDF-Berechnungen finden sich im Vektor-Space-Modell von Gerald Saltons SMART-Information-Retrieval-System Anfang der 70er-Jahre. Auf diesem Modell basieren auch heute noch die Kernfunktionen der meisten Suchmaschinen. Da es damals nur wenige Textdokumente in digitaler Form gab und das World Wide Web in weiter Ferne lag, wurde in Testcollections gesucht, die z. B. aus 400 Artikeln des Nachrichtendienstes Reuters, 600 Artikeln der Times oder 2.000 Abstracts wissenschaftlicher Artikel bestanden. Die damaligen Suchmaschinen wurden von Bibliothekswissenschaftlern entwickelt. Damals waren Bibliothekare noch ein viel wichtigerer Bestandteil der wissenschaftlichen Forschung. Sie halfen Wissenschaftlern, in Universitätsbibliotheken relevante Literatur für die jeweiligen Fragestellungen der Wissenschaft zu finden, und hatten den Überblick über wissenschaftliche Zeitschriften und Literatur in Buchform. Von ihren Fähigkeiten hing es oft ab, ob ein Wissenschaftler relevante Literatur zu seinem Thema fand oder nicht. Bibliothekare sind für die Verschlagwortung neu eintreffender Bücher und Zeitschriftenartikel zuständig. Nicht verwunderlich, dass sie sich eine Maschine wünschten, die ihnen auf Knopfdruck die wich-

DER AUTOR

Dr. Arne-Christian Sigge ist im Vorstand der content.de AG. Der Marktplatz für die Erstellung von Text-Content verfügt über ein eigenes Tool für WDF*IDF-Analysen.

tigsten Terme eines Dokumentes heraussucht. Zu diesem Zweck wurden Konzepte wie die WDF*IDF-Termgewichtung erschaffen.

WDF*IDF-Analysen dienten dazu, die Bedeutung von Termen in einem Dokument zu bestimmen, um so anhand der wichtigsten Terme den Inhalt des Dokuments für eine Suchmaschine erkennbar und greifbar zu machen. Sie waren ganz klar nicht als Konstruktionsvorschrift für die Texterstellung gedacht.

Warum graben SEOs solche alten Kamellen wieder aus?

Findige SEOs kamen auf die Idee, WDF*IDF-Berechnungen als Grundlage für Textkonstruktionsvorschriften zu verwenden, um Texte zu produzieren, die aus Sicht einer Suchmaschine so ähnlich aussehen wie die Texte, die zu einer bestimmten Keywordphrase gut ranken. Die Idee dahinter:

1. Finde Seiten, die zu einer vorgegebenen Keywordphrase gut ranken.
2. Bestimme die WDF*IDF-Werte der in allen Dokumenten verwendeten Terme je Dokument.
3. Bestimme die durchschnittlichen und maximalen WDF*IDF-Werte je Term und sortiere sie absteigend nach den durchschnittlichen WDF*IDF-Werten je Term.
4. Wähle Terme aus, die anhand der Aufreihung wichtig und sinnvoll erscheinen, und merke sie für einen eigenen Text vor.
5. Finde Terme, die in der Kurve bisher nicht deutlich in Erscheinung treten, um eine individuelle Nische zu besetzen.
6. Nimm den Bereich zwischen dem Durchschnitts- und dem Maximalwert als Zielraum für die Terme in einem eigenen Text.
7. Schreibe einen eigenen guten Text mit den wichtigen Termen, deren WDF*IDF-Werte sich im Idealfall zwischen die beiden Kurven einfügen und einen neuen Termraum schaffen, mit dem sie die bestehenden Texte ausstechen.

So weit keine schlechte Idee, um sich in einer halbwegs natürlichen Umgebung den Mitstreitern möglichst ebenbürtig, aber auch nicht übertrieben anzupassen, und in ihren Schwachstellen einen Ansatzpunkt zu suchen, sich selbst darüber hinaus zu profilieren.

Worin unterscheiden sich aktuelle WDF*IDF-Tools von den damaligen Verfahren?

Die bei der Entwicklung von WDF*IDF-Techniken verwendeten Testcollections bestanden aus journalistischen Texten angesehener Zeitungen und Agenturen sowie aus den Abstracts wissenschaftlicher Arbeiten. Gerade bei Letzteren, die größtenteils einen Peer-Review*-Prozess durchlaufen haben, kann man davon ausgehen, dass hier jedes Wort genau durchdacht und mehrfach abgewogen war. Ideale Ausgangsbedingungen für eine Suchmaschine. Die heutigen WDF*IDF-Tools verwenden als Dokumentkorpus eine Handvoll Webseiteninhalte, meist die ersten 10, 20 oder sogar 30 Trefferseiten zu dem vorgegebenen Suchbegriff. Jedes Analyse-Tool muss die Texte der für die Analyse zugrunde liegenden Webseiten extrahieren und speichern. Hier fängt der Spaß schon an: Viele Webseiten enthalten kein sauberes HTML, schlampen bei der Wahl von Zeichensätzen, den http-Headern, schummeln bei Trennungen und Zeilenumbrüchen, nutzen Javascript, um Texte nachzuladen, oder CSS, um sie auszublenden, und umgeben den eigentlichen Text mit einen Wust von Standardbausteinen wie Navigationselementen, Kopf- und Fußzeilen und Ähnlichem. Kurzum: Einen sauberen Text zu extrahieren, ist gar nicht so einfach und erfordert viel Fleißarbeit von Programmieren. Auch wenn hier vorbildlich gearbeitet wird, reichen die Ergebnisse in vielen Fällen nicht an die Qualität der damaligen Testcollections heran.

Neben der etwas schmutzigeren Datenbasis steht der Anwender vor einem ganz anderen Problem. Eine Testkollektion weist meist eine schöne inhaltliche Streuung zwischen den Dokumenten auf, da sie aus Nachrichten aus den Bereichen Politik, Kultur, Sport und Wirtschaft bestehen oder Forschungsbeiträge aus den verschiedensten Wissenschaftsdisziplinen anreißen.

Ein Dokumentkorpus, das hingegen nur aus 10, 20 oder 30 Treffern zu einem bestimmten Suchbegriff besteht, streut inhaltlich natürlich deutlich weniger. Per definitionem besteht es nur aus Dokumenten, die sich thematisch um ein einziges Thema, z. B. Baufinanzierungen, drehen. Schon stehen wir vor einem weiteren Problem: WDF*IDF-Werte errechnen sich bekanntlich aus zwei Komponenten. Der WDF-Anteil (Within Document Frequency) gibt dabei die Bedeutung eines Terms in einem Dokument an. Ganz grob gesagt: Das ist das, was der gemeine SEO früher mit der Keyworddichte erschlug, mit ein wenig Logarithmus drum herum. Die IDF-Komponente (Inverse Document Frequency) bildet die Bedeutung eines Wortes für ein Dokument ab, bezogen auf das ganze Dokumentkorpus, also alle betrachteten Dokumente. Die Idee dahinter: Ein Wort, das in allen Dokumenten auftaucht, kann keine große Bedeutung haben. Klassisches Beispiel sind hier Stoppwörter wie und oder Begriffe wie Forschung, die in vielen Abstracts eines wissenschaftlichen Papers auftauchen. Ein Term wie Herzklappenfehler, der nur in zwei Dokumenten erscheint, stellt einen deutlichen Unterschied zu den anderen Dokumenten dar und ist für die Kennzeichnung dieser beiden Dokumente von größerer Bedeutung als der Term Forschung.

Hier fällt uns dann auch gleich die

*siehe Online-Glossar unter www.websiteboosting.com

Einschränkung des Dokumentkorpus auf die Treffermenge zu einem bestimmten Keyword auf die Füße. Ist es nicht gerade eine Eigenschaft einer solchen Treffermenge, dass nicht nur das gesuchte Keyword, auf das der SEO seinen Text WDF*IDF-optimieren möchte, sondern auch viele semantisch naheliegende Terme in praktisch allen Dokumenten der Treffermenge vorkommen? Eben! Daher wird der Einfluss der IDF-Komponente von modernen WDF*IDF-Tools auch stark gedämpft.

Diese Dämpfung erfolgt meist in dem dritten Faktor, der mit modernen WDF*IDF-Analyse-Tools eingeführt wurde. Karl Kratz und onpage.org haben ihr p, SEOlyze errechnet statt eines WDF*IDF-Wertes einen EPS-KF-Wert und content.de bezeichnet die Termgewichtung als CRR-Wert (Content Relevance Rank). Jeder Hersteller würzt seine WDF*IDF-Berechnung durch sein persönliches Betriebsgeheimnis, in das die individuellen Google-Verschwörungstheorien, z. B. über die Bedeutung von Termpositionen im Text, einfließen. Im Endeffekt kommen allerdings alle Tools zu sehr ähnlichen Ergebnissen.

Wo liegen denn dann die Probleme bei der Anwendung?

Was verbinden viele Männer spontan mit Frauen? Genau: Schuhe, Kleider, Schminke und Oberweite. Wenn ein Mann als Frau erkannt werden will, trägt er demnach ein schrilles Kleid über einem großzügig ausgestopften BH, schminkt sich auffällig und trägt markante Schuhe mit hohen Absätzen. Willkommen in der Travestieshow. Nichts anderes passiert bei vielen WDF*IDF-Optimierungen von Texten. Wenn man nicht der ist, der zu sein man vorgeben möchte, müssen bestimmte Merkmale geändert, verstärkt oder vertuscht werden. Willkommen in der Travestieshow für Texte.

Fehler Nr. 1: Der Glaube, ein WDF*IDF-Tool mache einen schlechten Text zu einem guten Text

Konkret bedeutet das: Ist die Ausgangsbasis weit vom Zielraum entfernt, wirken die Anpassungen irgendwann auffällig und überzogen. Eine WDF*IDF-Optimierung macht keinen schlechten oder thematisch unpassenden Text zu einem guten Text mit Mehrwert für den Leser, ebenso wenig, wie viel Schminke einen Mann zur Frau macht.

Ein fokussierter Text, der alle relevanten Aspekte eines Themas erfasst und dem Leser durch neue Aspekte einen Mehrwehrt gegenüber anderen Texten verschafft, ist ein guter Text, der auch gut ranken wird. Hohe Verweildauer und natürliche Backlinks werden diesen Text auszeichnen. Ein Text, der wie eine überschminkte Tunte daherkommt, wird eventuell kurzfristig gut ranken, dann aber schnell wieder in der Versenkung verschwinden.

Fehler Nr. 2: Ein Klon sticht unter Klonen nicht hervor

Nächstes Problem: Was passiert, wenn alle Texte in den Suchergebnissen plötzlich WDF*IDF-optimiert sind? Gar nicht so unwahrscheinlich bei hart umkämpften Keywords. Plötzlich richtet man sich an einem Schönheitsideal aus, das sich gar nicht mehr an dem ursprünglichen Leitbild orientiert, sondern nur noch an den Nachahmern. Es sollte übereifrigen SEOs zu denken geben, dass, sobald alle Teenager mit Madonnas aktueller Frisur herumliefen, sich Madonna selbst schon wieder ein neues Haarstyling zugelegt hatte. Wer einem Trendsetter nacheifert, hat viele Nebenbuhler.

Fehler Nr. 3: Das WDF*IDF-Tool wird überschätzt und Inputdaten werden nicht ausreichend reflektiert

Mathematische Modelle reagieren oftmals sensibel, wenn man an den Inputwerten wackelt, die sich mit großem Hebel auf das Ergebnis auswirken. WDF*IDF-Analysen reagieren streckenweise auch empfindlich auf verschmutzten Dateninput. Das Gefährliche daran ist, dass sie dabei nicht einfach eine rote Warnlampe leuchten lassen und jegliche Ergebnisausgabe verhindern. Viel schlimmer: Sie versorgen den Autor mit falschen Daten und schicken ihn in die falsche Richtung. Wird hier blind den Anweisungen vertraut, verläuft man sich schnell auf dem Weg zum Ziel.

Der Ansatz, einen Text zu konstruieren, der sich nach den ermittelten Termen und WDF*IDF-Werten des zugrunde liegenden Suchergebnisses richtet, basiert auf der Annahme, dass das Suchergebnis aus guten Texten, die für das Keyword relevant sind, besteht. Leider ist das in vielen Fällen aber nicht gegeben, was gern übersehen wird. Nach dem ersten Rechenergebnis eines WDF*IDF-Analyse-Tools ist es daher Pflicht, die zugrunde liegende Treffermenge genau unter die Lupe zu nehmen. Dann zeigt sich schnell, ob die angezeigte Kurve überhaupt relevant für das angedachte Vorhaben ist und welche Werte und Terme des Analyseergebnisses ignoriert werden müssen. Hier ist der Mensch mit eingeschaltetem Hirn auf Betriebstemperatur gefragt.

Suchergebnisse ranken nicht wegen des Textes

Der Text ist nur einer von vielen Rankingfaktoren. Dieser immer wieder heruntergebetete Warnhinweis wird oft ignoriert. Manche Seiten ranken aufgrund guter Backlinks oder der Reputation der Domain. Ein gutes Beispiel dafür ist der Bausparrechner auf n-tv.de. Auf dieser Seite gibt es keinerlei Text zum Thema Vergleich von Bausparverträgen. Textuell ist neben den Header-, Footer- und Navigationselementen nur ein rechtlicher Disclaimer auf der Seite

zu finden. Trotzdem platziert sich die Seite unter den Top-10-Suchergebnissen zum Keyword Bauspartarife. So spülen solche Seiten auch gern mal Begriffe wie n-tv in die Termlisten. Termen von solchen Webseiten wird ein Einfluss auf das Ranking zugerechnet, der eben nicht in den Termen, sondern in anderen Faktoren begründet ist.

Suchergebnisse beinhalten keinen natürlichen Text

Ebenso unbrauchbar für WDF*IDF-Analysen sind Seiten, die keinen natürlichen Text enthalten. Zu den Top-10-Treffern der Keywordphrase *BMW-Jahreswagen* gehören Ergebnisseiten der Gebrauchtwagenbörse von bmw.de und der großen Autoportale *mobile.de* sowie *autoscout24.de*. Auf allen Seiten werden 20–30 aktuelle Angebote samt Sonderausstattung vorgestellt. Begriffe wie *Navi, Lederausstattung* und der Markenname *BMW* verzerren hier stark das Bild, das man von einem Text über *BMW-Jahreswagen* bekommen sollte.

Störend können sich auch Kommentarfunktionen auf Webseiten auswirken. Je nach Qualität der Seite und Besucher oder thematischer Ausrichtung sammeln sich unter dem eigentlichen Text Lobeshymnen, Beschimpfungen oder zahlreiche Posts, die mit dem Satz *„Ich hab da auch mal ne Frage …“* beginnen.

Suchergebnisse variieren thematisch

Zu einigen Keywords gibt es auf den ersten SERPs* der Suchmaschinen thematisch stark variierende Treffer. Sucht man nach *NDA*, findet man nicht nur Seiten zu *None Disclosure Agreements*, sondern auch gut positionierte Seiten der *Nuclear Decommissioning Authority*, der *National Development Agency*, der *National Dart Associaten* oder auch der *National Dance Association*.

Ein Analyse-Tool hat auch kein Gefühl dafür, ob es sich um Suchergebnisse zu einem hart umkämpften Keyword wie z. B. *private Krankenversicherung* oder um ein von SEOs bisher kaum beachtetes Nischenkeyword wie *Koi-Ablaichbürste* handelt, für das es kaum Treffer gibt. So kann es sein, dass das Dokumentkorpus nur aus SEO-optimierten Texten besteht oder dass nur wenige, ganz natürlich entstandene Seiten sich tatsächlich mit dem Thema befassen. Ebenso wenig kann das Tool erkennen, ob alle betrachteten Texte überhaupt für das Thema relevant sind.

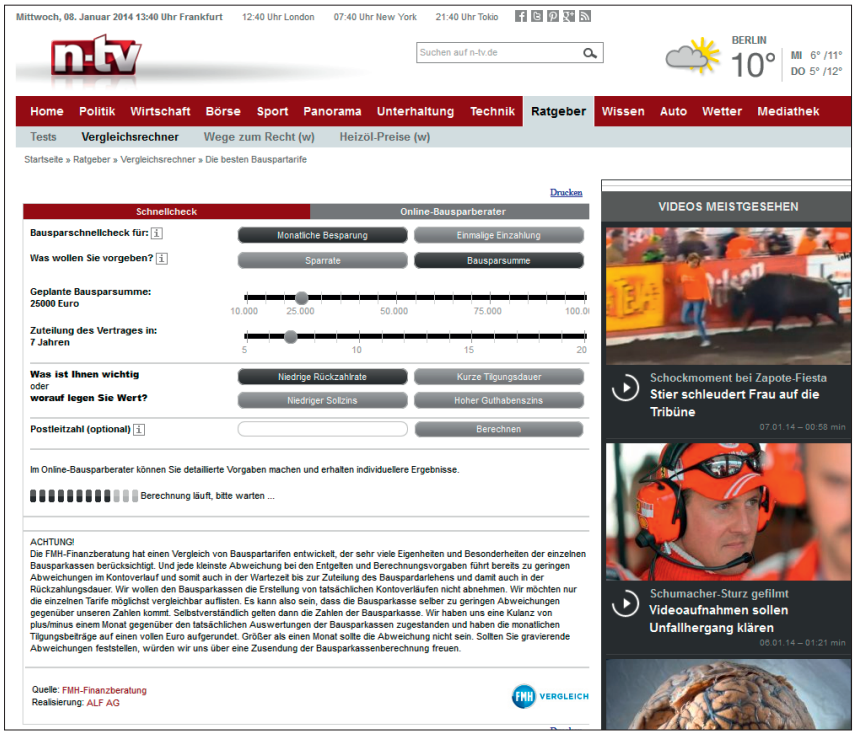

Abb. 1: Die Seite von n-tv rankt nicht aufgrund des Textes zum Thema Bauspartarife unter den Top-10-Treffern; ein Teil der Seite wird per Javascript nachgeladen, enthält jedoch auch keinen relevanten Text

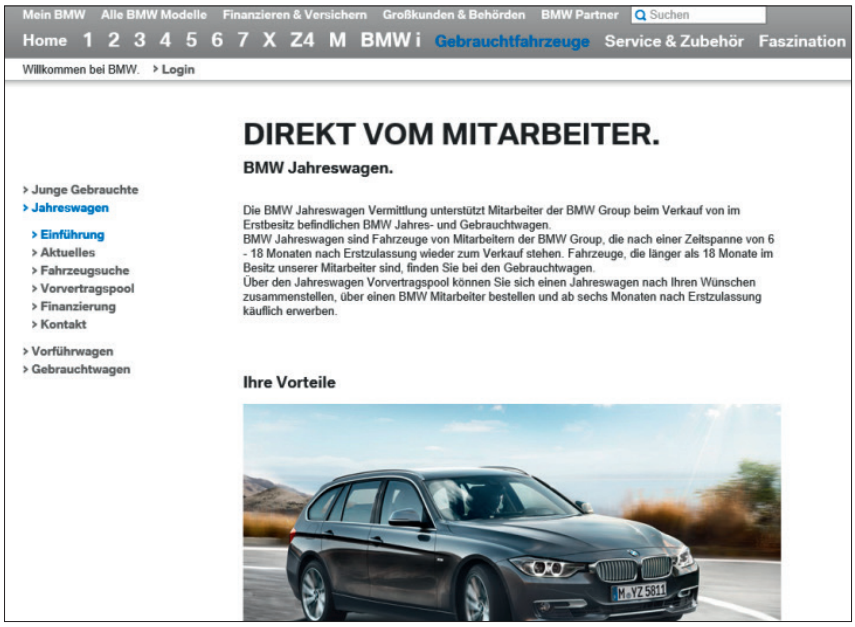

Abb. 2: Eine der wenigen Trefferseiten, die tatsächlich etwas Text zum Thema BMW-Jahreswagen enthalten

Terme im Suchergebnis müssen themenbezogen entwertet werden

Je nach Thematik müssen bestimmte Terme hinterfragt und aussortiert werden. Der Klassiker ist hier der Term

*siehe Online-Glossar unter www.websiteboosting.com

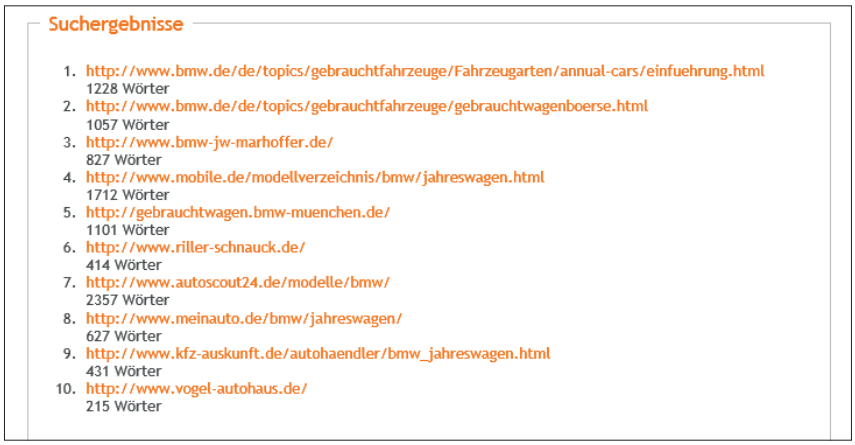

Abb. 3: Vor dem Blick auf die Analyseergebnisse ist dringend die Datengrundlage unter die Lupe zu nehmen. Welche Seiten liefern den Input für die WDF*IDF-Kurven?

Abb. 4: Nach einem kurzen Text über BMW-Jahreswagen folgen auf mobile.de zahlreiche Fahrzeugangebote mit Ausstattungsdetails

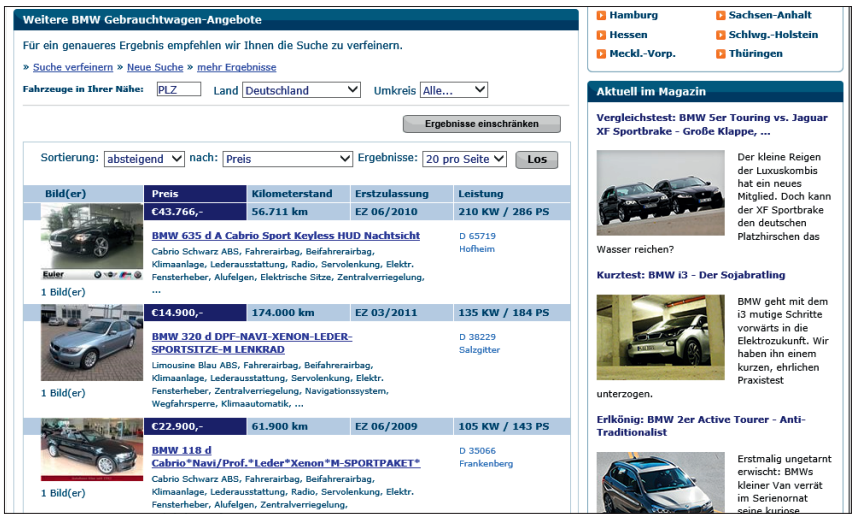

Abb. 5: Auch auf Autoscout24.de besteht der Großteil des Textes aus der Aufzählung von Ausstattungsdetails

Impressum. Fast auf jeder Webseite ist dieser Term mindestens einmal zu finden, weil der Seitenbetreiber andernfalls abgemahnt werden kann. Grundsätzlich kann er daher auch ignoriert werden. Ausnahmen gibt es sicherlich in Texten über *Abmahnfallen für Online-Shop-Betreiber*. Der Term *suchen* kann aus den Beschriftungen von Suchformularen in die Liste der bedeutenden Terme gerutscht sein. Beim Thema Immobilien kann er auch eine wichtige inhaltliche Bedeutung haben, schließlich suchen Besucher von Immobilienportalen Grundstücke, Wohnungen oder Häuser. Gleiches gilt für Jahreszahlen. Auch hier ist fallweise zu hinterfragen, ob ihr Auftreten thematisch von Bedeutung ist.

Bei anderen Suchanfragen muss man den zeitlichen Kontext der Analyse mit einbeziehen. Zum Thema *Fußball-WM* findet man aktuell viele Texte mit dem Term *Brasilien* und der Jahreszahl 2014. Das kann nach der WM schon wieder ganz anders aussehen. Bei einer richtigen Gewichtung bestimmter Terme muss man auch berücksichtigen, dass es zwischen den Weltmeisterschaften in Deutschland, Südafrika und Brasilien eine erhebliche Entwicklung beim Textwachstum und der SEO-Aktivität im Internet gab. War die WM 2006 in Deutschland weniger wichtig als die WM 2010 in Südafrika oder lassen sich Analyse-Tools hier durch Umfang und Aktualität der Datenbasis täuschen?

Wie sollte eine WDF*IDF-Analyse bei der Content-Erstellung eingesetzt werden?

Sind WDF*IDF-Analyse-Tools demnach unbrauchbar? Keinesfalls, sie müssen nur mit Verstand eingesetzt werden. Es darf nicht der Irrglaube herrschen, mit einem WDF*IDF-Zauberknopf mache man einen schlechten Text zu einem guten Text. Ein Bildchen mit zwei Kurven und einem Schieberegler darf nicht die Ausgangsbasis für einen Text sein. Bevor

man überhaupt einen Blick auf die Analyseergebnisse wirft, ist zwingend einer auf die als Grundlage der Analyse genutzten Webseiten nötig. Diese Seiten müssen unter den Gesichtspunkten natürlicher Text, andere Rankingfaktoren, Ergebnisumfeld etc. unter die Lupe genommen werden. Hier sollten SEO-Fachmann und Autor gemeinsam die Analyseergebnisse betrachten. Dabei müssen Terme der Ergebnisliste und Webseiten des Dokumentkorpus aussortiert werden. Danach muss abgewogen werden, inwieweit die verbliebenen WDF*IDF-Berechnungen noch brauchbar sind. Schließlich verändert man durch Ausschlüsse die Grundgesamtheit. Ein WDF*IDF-optimierter Text ist nur so gut wie der SEO, der die Analyseergebnisse vorher unter die Lupe genommen und daraus einen Textauftrag für den Autor erstellt hat. WDF*IDF-Analysen eignen sich definitiv nicht für eine automatisierte Massentextproduktion auf Knopfdruck. Sie sind erst der Anfang einer weiterführenden manuellen Analyse.

Ein Autor steht weiterhin vor der Aufgabe, einen fokussierten Text mit Mehrwert für den Leser zu schreiben, der alle Aspekte des Themas angemessen berücksichtigt. Genau dafür sollte er die WDF*IDF-Analyse nutzen: vorab als Rechercheinstrument, das ihm aufzeigt, welche Themen relevant sind, mit welchen Termen diese gewöhnlich verbalisiert werden und wie einzelne Aspekte zu gewichten sind.

Die Kurven einer WDF*IDF-Analyse unterstützen ihn im Anschluss an die Texterstellung ggf. dabei, seinen persönlichen Formulierungsstil dem allgemeinen Sprachduktus zu dem jeweiligen Thema anzupassen.

Wer einfach versucht, irgendeinen mittelprächtigen Text zwischen zwei Kurven zu prügeln, deren Hintergrund nicht hinterfragt wird, ist auf dem Holzweg – und der führt bekanntlich nicht auf die vorderen Plätze in den Suchergebnissen. ¶

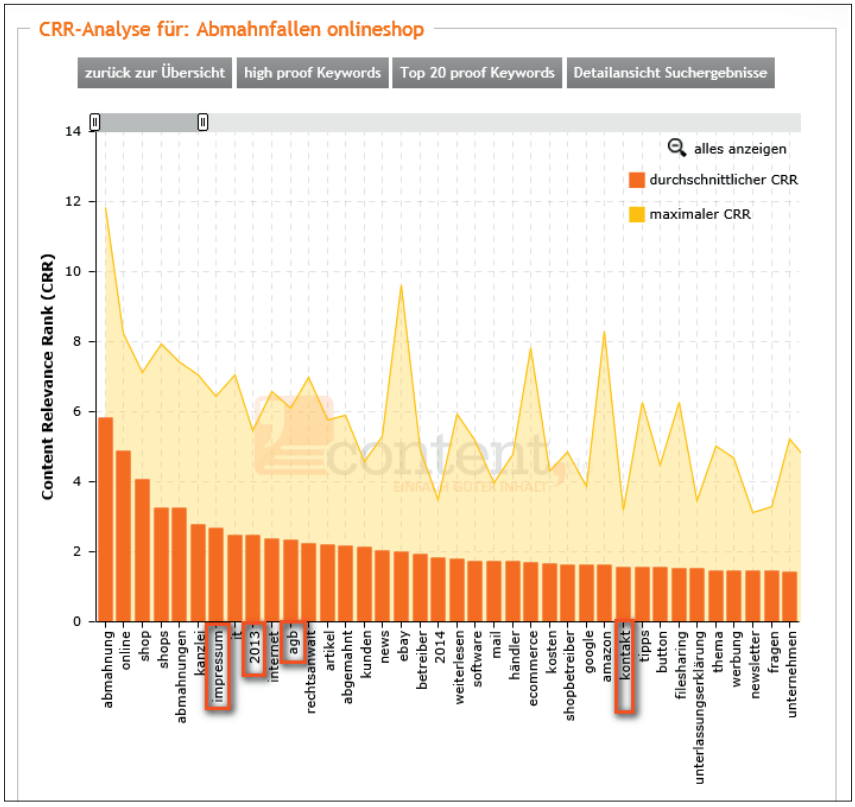

Abb. 6: Bestimmte Begriffe müssen themenbezogen betrachtet werden; die Klassiker AGB, Impressum und Kontakt können oft ignoriert werden, beim Thema Abmahnfallen Online-Shop sind sie allerdings zu berücksichtigen

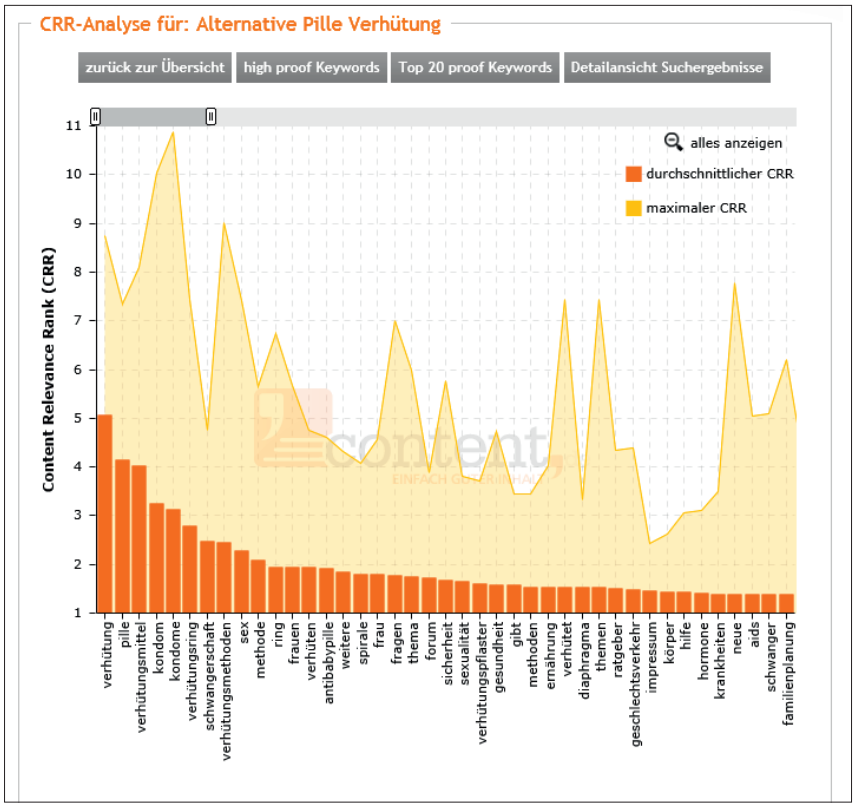

Abb. 7: Eine WDF*IDF-Analyse dient im ersten Schritt als Recherchehilfe für den Autor; sie zeigt auf, welche Aspekte zu beachten sind und wie stark sie bei den konkurrierenden Texten im Fokus stehen