Mario Fischer

Serie: SEO für Einsteiger - Teil 3

Ich will das Steak - nicht die Pfanne!

Wie wertet Google die Inhalte auf einer Seite?

Aufgrund vieler Anfragen und Anregungen unserer Leser starteten wir in Ausgabe 68 erneut eine kleine Serie mit SEO-Basics für Einsteiger und motiviert Fortgeschrittene. In den bisherigen Teilen ging es um den Title, die Description und die korrekte Anwendung von Überschriften (H1, H2 usw.).

Viele Websitebetreiber und auch nicht wenige Webagenturen sind der Meinung, dass Google den Text auf einer Webseite für das Ranking heranzieht. Doch das stimmt nur zum Teil. Während fast alle Tools bei Textanalysen auf den kompletten Text zugreifen und diesen auswerten, grapscht sich Google in der Regel gezielt den sog. „Primary Content“ für die Gewichtung des Rankings. Der übrige Text wird nicht oder nur in geringem Maße berücksichtigt.

DER AUTOR

Mario Fischer ist Herausgeber und Chefredakteur der Website Boosting und seit der ersten Stunde des Webs von Optimierungsmöglichkeiten fasziniert. Er berät namhafte Unternehmen aller Größen und Branchen und lehrt im von ihm gegründeten Studiengang E-Commerce an der Hochschule für angewandte Wissenschaften in Würzburg.

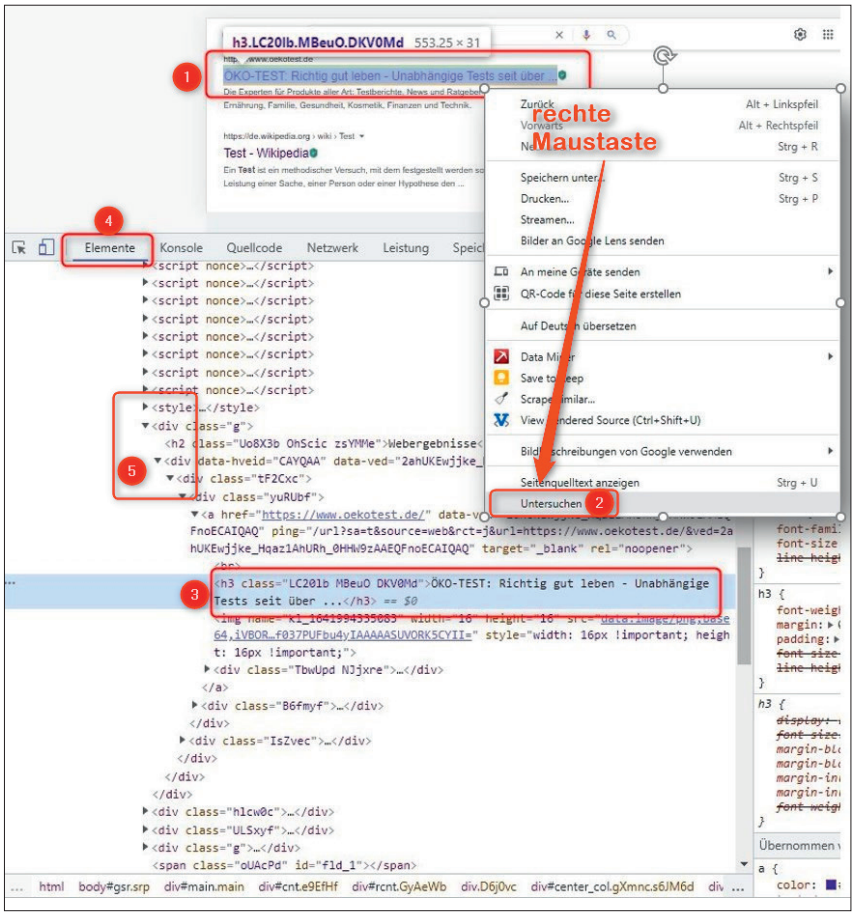

Abb. 1: So sieht der DOM eines Google-Suchergebnisses aus

Abb. 2: Der DOM ist hierarchisch aufgebaut

Veganer mögen die Überschrift verzeihen, sie soll nur plakativ und nicht wertend darstellen, was Googles Absicht bezüglich der textlichen Rankingsignale ist. Schaut man sich beliebige Webseiten an, fällt auf, dass fast alle ein Stück Steak haben, also den gewünschten Text, wegen dem man die Seite besucht hat. Und sie haben auch eine sog. Boilerplate, frei interpretiert: die Pfanne. Das ist der umgebende Text, der beim Navigieren hilft, erklärt, welche Website das ist, andere Inhalte anteasert und sonstigen Text enthält, der nicht Kern dieser Seitenbotschaft ist. Mit Boilerplate bezeichnet man einen häufig verwendeten Textbaustein, Standardformulierungen, Mustertexte und Weiteres, was eben mehrmals verwendet wird, weil es zusätzlich tatsächlich oder vermeintlich notwendig ist. Am besten kann der sog. Footer, also die Fußzeile auf Webseiten, für das herhalten, was man als Boilerplate bezeichnet. Der ist in der Regel auf jeder Seite einer Domain gleich. Macht es also Sinn, solche Texte, die auf allen oder zumindest vielen Seiten einer Domain enthalten sind, für das Ranking einer individuellen Seite heranzuziehen? Wohl eher nicht.

Also muss man den Text finden, der die Seite ausmacht, der sie „unique“, also einzigartig macht. Das klingt allerdings einfacher, als man sich das unbedarft vorstellt. Tatsächlich ist es eines der aufwendigsten Verfahren, den Primary Content automatisiert zu finden! Dazu braucht es enormen technischen Aufwand, wie die folgende und durchaus lehrreiche Erklärung zeigt.

Wie ist eine HTML-Seite *wirklich* strukturiert?

Ganz einfach, werden Sie vielleicht denken. Da gibt es Überschriften, die Textblöcke gliedern. Für unser menschliches Auge stimmt das natürlich. Aber Maschinen haben keine solchen „intelligenten“ Augen, sie sehen nur

programmierten Code. In der Tat haben HTML-Dokumente eine innere Struktur, den sog. DOM – das Document Object Model. Genau genommen ist es gar kein „Modell“, sondern eher eine Art Programmierschnittstelle. Aber das führt uns zu weit und ist für das Verständnis nicht notwendig. Dieser „Objektbaum“ ist hierarchisch gegliedert und enthält jedes Element, das für den Aufbau und die Darstellung der Webseite verantwortlich ist.

Was der DOM ist und wie er aussieht, kann man leicht visualisieren. Man öffnet z. B. den Chromebrowser, ruft eine Webseite auf und markiert mit der Maus einen Bereich auf der Seite, also eine Überschrift, einen Textblock oder Ähnliches, wie Ziffer 1 in Abbildung 1 zeigt. Die rechte Maustaste bringt dann ganz unten „Untersuchen“ als Auswahlpunkt hervor (Ziffer 2). Klickt man diesen an, erscheint die sog. Entwicklerkonsole unten oder rechts (je nach Voreinstellung) im Browser. Oben sieht man also noch einen Teil der Webseite und unten die Codierung bzw. den DOM. Dieser ist bereits genau an der Stelle aufgeklappt (Ziffer 3), wo sich das eben auf der Seite markierte Element befindet. Die Entwicklerkonsole hat mehrere Menüpunkte, unter „Elemente“ (Ziffer 4) findet man den Objektbaum der Seite.

Bei Ziffer 5 sieht man recht deutlich die Dreiecke, die nach rechts oder nach unten zeigen. Dort lässt sich der Baum auf und zu bzw. tiefer klicken. Probieren Sie das ruhig einmal aus. Das meiste ist Programmierkram, aber irgendwo in diesem Baum verbergen sich tatsächlich alle Elemente wie Texte, Bilder, Links etc. einer Webseite. Über die Methode „markieren und rechte Maustaste“ plus „Untersuchen“ findet man diese automatisch und muss nicht danach suchen.

Es ist leicht einzusehen, dass Textblöcke je nach CMS, Shopsystem oder gar mittels manueller Seitenprogrammierung überall in diesem DOM verortet sein können. Ganz oben, unten, flacher oder tiefer in der Hierarchie liegend. Jedes Element hat also seine individuelle – im wahrsten Sinne individuelle! – „innere“ Adresse im Dokument. Die Adresse eines Dokuments im Web wird durch die URL festgelegt, wie z. B.:

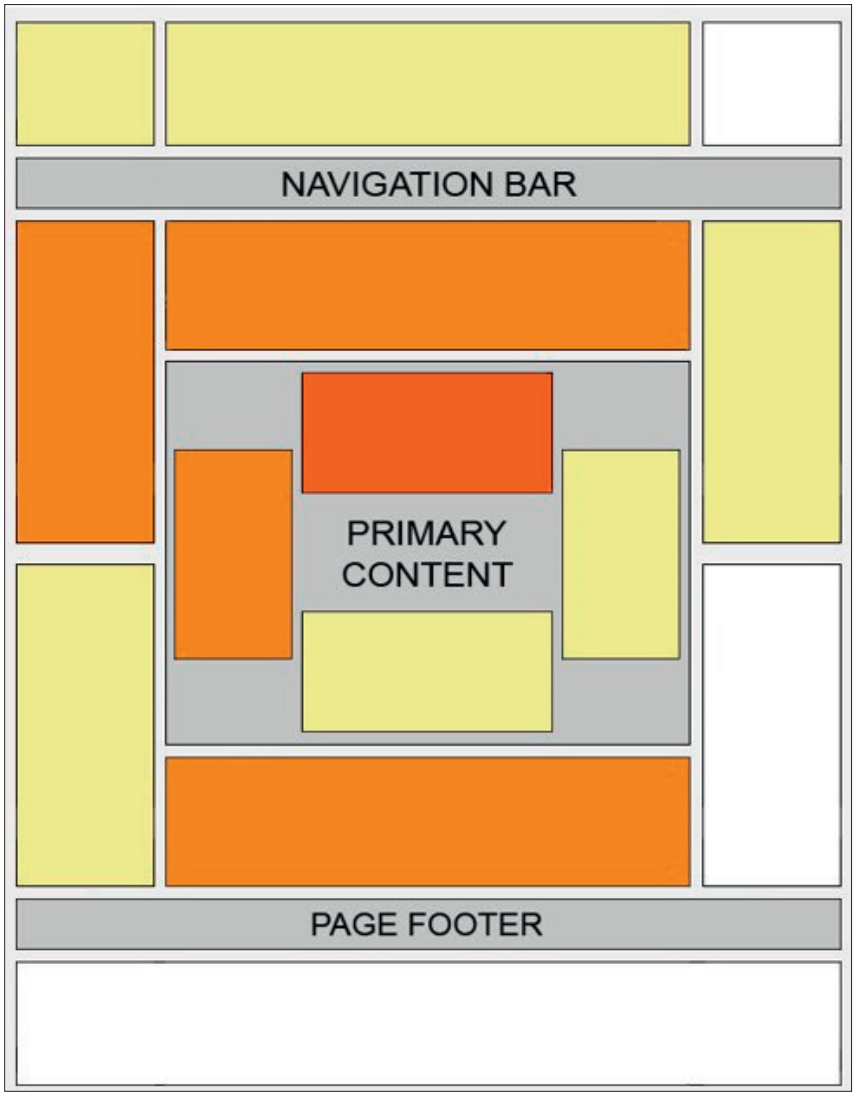

Abb. 3: Eine historische Abbildung von Google über den Primary Content *(Quelle: google.com)*

```
https://www.google.com/search?q=test&[...]
```

Und dort findet man dann als Elementadresse z. B.:

```
//*[@id="rso"]/div[3]/div/div[1]/a/h3
```

Wer den Begriff „Xpath“ schon einmal gehört hat, dies ist er. Was hier in einer Zeile dargestellt wird, muss man sich als Hierarchie vorstellen:

```
//*[@id="rso"]
  /div[3]
    /div
      /div[1]
        /a
          /h3
```

Das letzte Element ist /h3, also eine Überschrift der Ebene 3, und enthält im realen Beispiel den Text, der auf der Suchergebnisseite als solcher dargestellt wird (siehe Abbildung 1 Ziffer 3). Ändert man den Aufbau der Seite, ist es möglich, dass sich die DOM-Adresse ändert. Es ist also nicht so, dass man ein bestimmtes Textstück innerhalb einer Domain immer an derselben DOM-Adresse findet. Sobald ich das Template, also die benutzte Vorlage,

Abb. 4: Durch Differenzanalysen kann der einzigartige Textbereich identifiziert werden

ändere, ändert sich auch der DOM-Aufbau. Berücksichtigt man nun, dass fast alle Domains mit eigenen Vorlagen arbeiten bzw. selbst Standardvorlagen oft modifiziert werden, wird deutlich, dass eine Suchmaschine den wichtigen Primary Content nicht einfach finden bzw. adressieren kann.

Menschen erkennen den Kern einer Seite schnell, Maschinen nicht

Wir Menschen haben damit kein Problem, ein Blick genügt. Früher war auch Google nicht dazu in der Lage und wertete alle auf einer Seite gefundenen Texte. Später konnte man dann differenzieren, was in Überschriften (Hx) steht, und das stärker als Wörter im Fließtext berücksichtigen. Im Lauf der Zeit wurde die Erkennung immer weiter vorangetrieben und jetzt ist man in der Lage, wie das menschliche Auge, den Primary Content zu erkennen – zumindest in den meisten Fällen. Vor über zehn Jahren zeigte Google für sein AdSense-Programm in der Hilfe, wie Seitenbereiche unterschieden werden können, wie Abbildung 3 zeigt (Quelle damals: *http://support.google.com/adsense/bin/answer.py?hl=de&ctx=as2&answer=1354747&rd=1*, mittlerweile gelöscht). Hier ging es allerdings um die optimale Platzierung von Anzeigen.

Doch wie macht Google das, wo diese Elemente doch wie oben beschrieben mindestens domainübergreifend andere Detailadressen in den Dokumenten haben? Natürlich könnte man einfach hergehen und die Anzahl Wörter in Textblöcken zählen, um zu ermitteln, wo vermutlich der „wichtige" Text steht. Aber das ist keine stabile Lösung, denn diese Textblöcke können von der Lage durchaus weiter auseinander sein, unterschiedliche Detailadressen haben, oder sie haben einzeln in Summe zu wenig Text, um sich vom restlichen Text in Footern oder Sidebars zu unterscheiden. Diese Lösung wäre also zu naiv.

Die Lösung ist schnell beschrieben, aber technisch unheimlich aufwendig, weil man dazu eine enorm hohe Rech-

nerpower und sehr viel Speicherplatz braucht. Man liest dazu alle DOM-Adressen bzw. Elemente aus, die eine gewisse Textmenge enthalten, und vergleicht sie seitenübergreifend innerhalb einer Domain. Sind diese Seiten sauber programmiert bzw. mit einheitlichen Vorlagen erstellt, findet man die Kopftexte, Footer, Navigationstexte etc. im DOM immer an der gleichen Stelle bzw. selbst bei wechselnden Stellen mit immer den gleichen oder ähnlichen Texten bzw. Textmustern. Sie wiederholen sich auf allen oder zumindest auf vielen Seiten.

Bei einigen Elementen mit genügendem Umfang ist der Text jedoch deutlich unterschiedlich von Seite zu Seite. DAS ist in der Regel die Primary Content – und er muss bei dieser Methode der Erkennung noch nicht einmal direkt hintereinander stehen. Diese Art der „Differenzanalyse", also „Welche Elemente variieren textlich auf allen Seiten?", bringt mit hoher Wahrscheinlichkeit genau die Textabschnitte, warum ein Besucher die Seite aufruft bzw. liest. DAS ist der Text, der für das Ranking herangezogen werden kann und der nicht mehr verrauscht bzw. „verschmutzt" wird durch Wörter, die nicht zu diesem Kern gehören!

Abbildung 4 zeigt dies schematisch auf. Die blau markierten Bereiche wiederholen sich auf allen Seiten, in der Regel auch inhaltlich. Der jeweils rot umrandete Bereich ist auf jeden Fall „unique", also einzigartig, und damit der Informationskern der jeweiligen Seite. Eine Ausnahme bildet die DOM-Adresse:

```
//*[@id="wp_author_widget-2"]/div[1]
```

Sie ist in der Abbildung hellblau hinterlegt und stellt den Autorenkasten dar. Wäre hier immer der gleiche Autor zu finden, würde man das eher zur Boilerplate zählen, variieren die Autoren und damit auch die Texte in dieser Box, ist das durchaus auch für die Einzelseite rankingrelevant. Im ersten Fall, überall der gleiche Name, wäre dieser eher der ganzen Domain zuzuordnen und eben nicht einer einzelnen Seite.

Abb. 5: Martin Splitt von Google erklärt, wie Topics und Sektionen einer Seite erkannte werden können *(Quelle: YouTube)*

Zur Topic-Erkennung ist ein enormer Rechenaufwand nötig

Man kann sich leicht vorstellen, wie aufwendig und ausgefuchst solche Differenzalgorithmen ausprogrammiert werden müssen und wie viel IT-Power man zur Verfügung haben müsste, um so etwas verlässlich zu ermitteln. Um die Boilerplate oder umgekehrt den Primary Content für eine einzelne Seite einigermaßen korrekt zu ermitteln, müsste man mindestens einen größeren Teil der Domain crawlen und auch jede dieser Seiten anschließend in DOM-Elemente zerlegen und diese für einen „Alles-mit-allem"-Vergleich abspeichern. Eine Webseite hat gut und gerne zwischen 1.000 und 2.000 Elemente, bei 1.000 Seiten wären das im Mittel in Summe 1,5 Mio. Elemente, die man einem solchen Ähnlichkeitsvergleich unterziehen müsste.

An dieser Stelle wird auch klar, warum textliche Analysen vieler SEO-Tools nicht so korrekt arbeiten (können), wie man sich das vielleicht ohne Kenntnis dieser Umstände vorstellen würde. Meistens packen diese jedes Wort einer Seite in einen Eimer und analysieren diese ohne Unterscheidung einer „Wichtigkeit", so wie es Google tut.

Google kann mittlerweile sehr viel mehr leisten, als man sich vielleicht üblicherweise vorstellen mag. Zumindest, wenn man den Aussagen bekannter und geschätzter Googler wie Martin Splitt vertrauen mag (YouTube-Video unter http://einfach.st/msplitt7):

> *„... So if you happen to have content on a page that is not related to the main topic of the rest of the content, we might not give it as much of a consideration as you think ..."*

Es kann bzw. wird also sehr wohl unterschieden zwischen Content, der relevant für das Topic einer Seite ist, und dafür irrelevantem Text. Das kann bei Tools, die diese Unterscheidung nicht treffen oder technisch bedingt nicht treffen können, durchaus zu Verzerrungen bei Analysen führen. Splitt erklärte weiterhin:

> *„... fundamentally we can read that from the content structure in HTML already and figure out so 'Oh! This looks like from all the natural language processing that we did on this entire text content here that we got, it looks like this is primarily about topic A, dog food.*

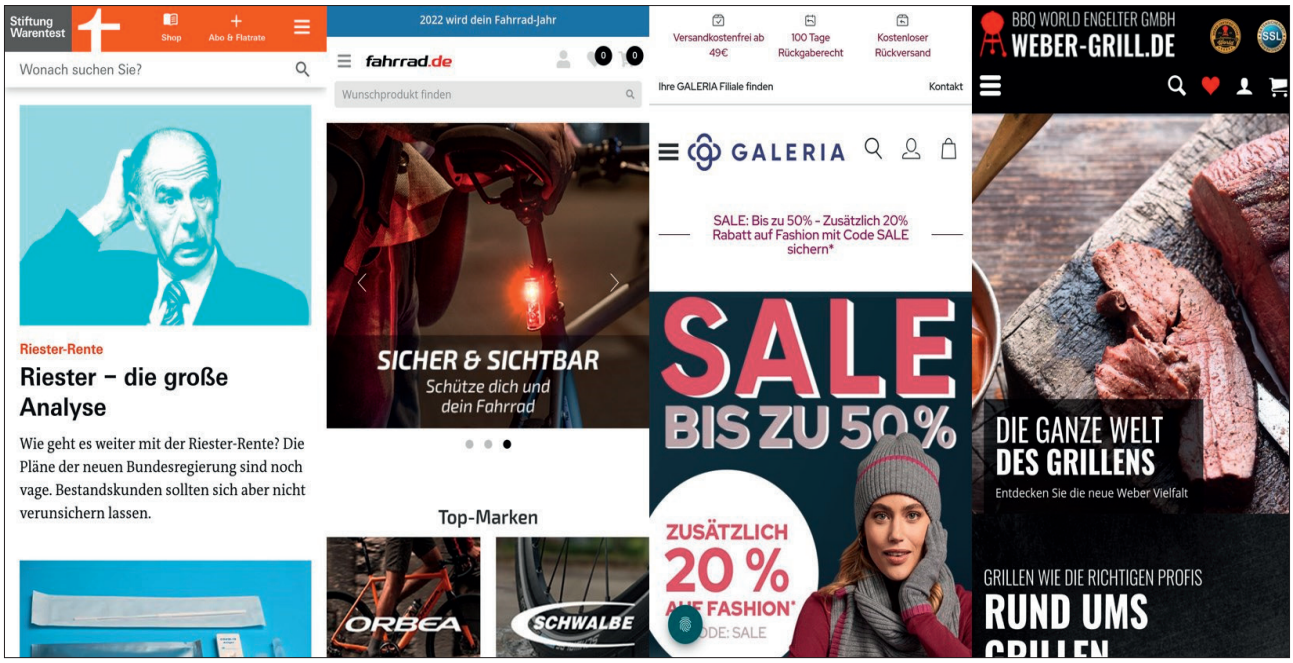

Abb. 6: Exemplarisch: „Above the fold“ auf dem Smartphone-Display

[…] And then there's this other thing here, which seems to be like links to related products but it's not really part of the centerpiece. It's not really main content here. This seems to be additional stuff. And then there's like a bunch of boilerplate or, 'Hey, we figured out that the menu looks pretty much the same on all these pages and lists. This looks pretty much like that menu that we have on all the other pages of this domain,' for instance, or we've seen this before. We don't even actually go by domain or like, 'Oh, this looks like a menu.'"

Der von ihm verwendete Begriff „Centerpiece Annotation" ist interessant, da er seitens Google so bisher nirgends auf den eigenen Webseiten erwähnt wurde – außer eben in der Aussage von Splitt. Wahrscheinlich soll es in etwa dem entsprechen, was Google ansonsten in Publikationen und Patenten als Primary Content bezeichnet, oder dem wichtigsten Teil innerhalb des Primary Contents – also einer noch feineren Unterteilung? Möglicherweise ist es aber auch ein interner Begriff, der zum ersten Mal so erwähnt wurde?

„… So I can probably say that we have a thing called the Centerpiece Annotation, for instance, and there's a few other annotations that we have where we look at the semantic content, as well as potentially the layout tree …"

Google kann also offenbar nicht nur wichtigen Text bzw. wichtige Textblöcke erkennen und stärker gewichten, sondern auch Links aus diesen Bereichen. Ebenso erklärte Splitt weiter, dass auch ein Vergleich der Textmenge zu einem Topic gegenüber der zu einem anderen Topic verwendet würde. Hat eine Seite z. B. 10.000 Wörter für Hundefutter und 3.000 Wörter zu Fahrrädern, wäre das wohl kein guter Content für das Topic Fahrräder, so sein Beispiel hierfür. Eine mögliche Schlussfolgerung wäre, dass die Seite damit eben nicht (gut) für Fahrräder ranken kann. Ob eine derartige „Verwässerung" von Topics auch dem Ranking für Hundefutter schaden könnte, wurde leider nicht erörtert, aber man darf es zumindest vermuten, weil es logisch ist. Jemandem eine Seite für eine Suche anzuzeigen, die sich nicht wirklich auf dieses Thema konzentriert, macht wohl immer dann Sinn, wenn es Alternativen gibt, die eindeutig das Contentsignal senden: „Ich bin eine (nahezu) 100%ige Fahrrad-Topic-Seite." Und derartige textreine Alternativen gibt es wohl mittlerweile viele im Web.

Beruhigend ist allerdings, dass bei einer sauberen Programmierung z. B. automatisch erstellte Contentteile mit Empfehlungen für andere Produkte mittlerweile gut erkannt und im Wesentlichen ignoriert werden. Sofern man nicht in gleichen Containern unterschiedliche Topics behandelt, dürfte das also wohl kein Problem darstellen.

Warum solche Überlegungen wichtig sind

Was wären also die Learnings aus diesen Überlegungen?

1. Zu viele unterschiedliche bzw. individuelle Templates könnten Google die Erkennung des Primary Contents erschweren. Stabilität im DOM bzw. bei dessen Elementen wäre in dieser Hinsicht daher durchaus positiv zu bewerten. Natürlich gibt es mit Sicherheit eine technische Fallback-Möglichkeit bei Google für die Inhaltserkennung und im „schlimmsten" Fall würde halt der gesamte Text für die Analyse herangezogen. Das wäre zwar kein Bein-

bruch, aber je genauer eine Analyse den unwichtigen, allgemeinen, sich wiederholenden Text von dem Kerntext einer Seite trennen kann, umso besser muss das logischerweise für das – textbasierte – Ranking dieser Seite sein. Damit ist der Anteil der vielen Algorithmen gemeint, die ein Scoring auf Textbasis durchführen. Wechselt die textliche Kerninformation ständig Lage und interne Adresse (XPath), wäre eine oben beschriebene Differenzanalyse sicherlich sehr viel schwieriger oder zumindest ungenauer bzw. mit mehr Fehlern behaftet durchzuführen.

2. Interne Links, die mitten im Primary Content platziert werden, sollten von Google deutlich stärker gewichtet werden als solche in Menüs, Sidebars oder gar Footern. Wer jetzt noch sprechende Ankertexte verwendet statt des nichtssagenden „mehr“ oder gar „hier klicken“, gibt den Links nicht nur den höheren Wert mit auf den Weg, sondern auch wichtige Keywords, die dem Ranking der Seite zugerechnet werden, auf die diese Links zeigen. Wichtige Links an diesen Stellen zu platzieren, scheint also eine durchaus lohnende Idee zu sein.
3. Bei den Ergebnissen von Textanalysen von SEO-Tools sollte man sich bewusst machen, ob das Tool den Primary Content maschinell bestimmen kann oder ob der gesamte Text einer Seite analysiert wird. Ersteres ist wirklich sehr aufwendig und die Tools, die dies können oder zumindest „versuchen“, nicht unbedingt billig. Preiswert ja, aber eben nicht für Spielgeld oder gar als kostenloses Tool zu haben. Auch hier ist es kein Beinbruch, wenn die Boilerplate mit einfließt, nur sollte man das dann eben immer im Kopf behalten, wenn man die Ergebnisse interpretiert bzw. verwendet.
4. Google erkennt durch Rendern des HTML-Codes, was „above the fold“ (über der Falz bzw. ohne Scrollen) liegt und gewichtet das „sofort Sichtbare“ offenbar deutlich stärker. Leider bleibt die folgende Frage offen: Das Konzept „above the fold“ kommt ja aus der Zeitungsbranche und bezeichnet den Bereich z. B. einer gefalteten Tageszeitung, der sichtbar ist, ohne die Zeitung aufzuklappen oder gar aufzuschlagen. Wegen dieses sichtbaren Bereichs entscheidet sich ein Käufer, die Zeitung in die Hand zu nehmen oder früher in den überall aufgestellten Zeitungskästen gegen Einwurf von Geld direkt „blind für den Rest“ zu kaufen. Was da steht, ist also enorm wichtig. Das hat man auf den Desktop übertragen, der ja einen ähnlichen Bildschirmaufbau hat, nämlich quer, wie eine Tageszeitung. Das Aufklappen der Zeitung entspricht dabei dem Nach-unten-Scrollen auf dem Bildschirm.
Aber Google betont mehrmals am Tag das „Mobile First“ und die Rankinganalyse ist schon seit Langem für die meisten Seiten auf diese mobile Anzeige, genauer auf die kleine Anzeige auf Smartphones umgestellt wurden. Wie steht es da eigentlich mit der Falz? Die meisten Webseiten zeigen auf Smartphones wenig oder oft sogar gar keinen Text, wie die willkürlich gewählten Beispiele in Abbildung 6 zeigen. Das Konzept „über der Falz“ macht also bei Smartphones wenig Sinn.
Oder kann das sogar eine enorme Chance sein, HIER einen kurzen, aber vernünftig überlegten Text anzuzeigen, während die oft designgesteuerten Mitbewerber Bilderwelten-Feuerwerke zünden, um den Besucher tatsächlich oder vermeintlich in Kauflaune zu bringen? ¶