
Яндекс
DAS YANDEX-L
RANKING-FAKT

..EAK: ITOREN

Mario Fischer

Zu Jahresbeginn wurden offenbar von einem Entwickler der russischen Suchmaschine Yandex neben vielen anderen Faktoren auch alle 1.922 Ranking-Faktoren geleakt. Für Rankinganalysen stellen die Aufstellungen ein wahres El Dorado dar. Einen so massiven Einblick in die Arbeit einer Suchmaschine gab es seit dem Bekanntwerden der AOL-Suchanfragendaten im Jahr 2006 nicht mehr.

Natürlich ist Yandex nicht Google und natürlich müssen die verwendeten Rankingsignale nicht übereinstimmen. Aller Wahrscheinlichkeit nach tun sie dies auch nicht – zumindest nicht genau oder so, dass man Regel eins zu eins für das Google-Ranking ableiten könnte.

Eine moderne Suchmaschine zu betreiben, lässt aber umgekehrt nicht allzu viel Spielraum für gute Signale. Mit anderen Worten kann man getrost davon ausgehen, dass Yandex keine komplett anderen Faktoren berücksichtigt als Google. Zudem haben in der Vergangenheit immer wieder Entwickler von Google bei Yandex angeheuert. Auch das ist kein Grund für einen direkten Know-how-Transfer. Vermuten darf man ihn vernünftigerweise aber eher wohl doch. Noch sind die Experten der Branche, allen voran Dan Taylor, Michael Kink und Alex Buraks, mit der Auswertung der vielen Daten beschäftigt und dass diese in kyrillischer Schrift und mit vielen Abkürzungen vorliegen, macht die Sache nicht einfacher. Was bisher bekannt ist, ist aber spannend genug, Ihnen hier eine kleine Zusammenfassung zu geben.

Dass die Daten echt sind, sagen einige Insider. Und ein weiteres Indiz spricht recht eindeutig dafür: Die Dateien im geleakten Code-Repository sind auf den 24. Februar 2022 datiert. Man darf getrost davon ausgehen, dass es kein Zufall ist, dass genau an diesem Datum der russische Einmarsch in die Ukraine stattfand.

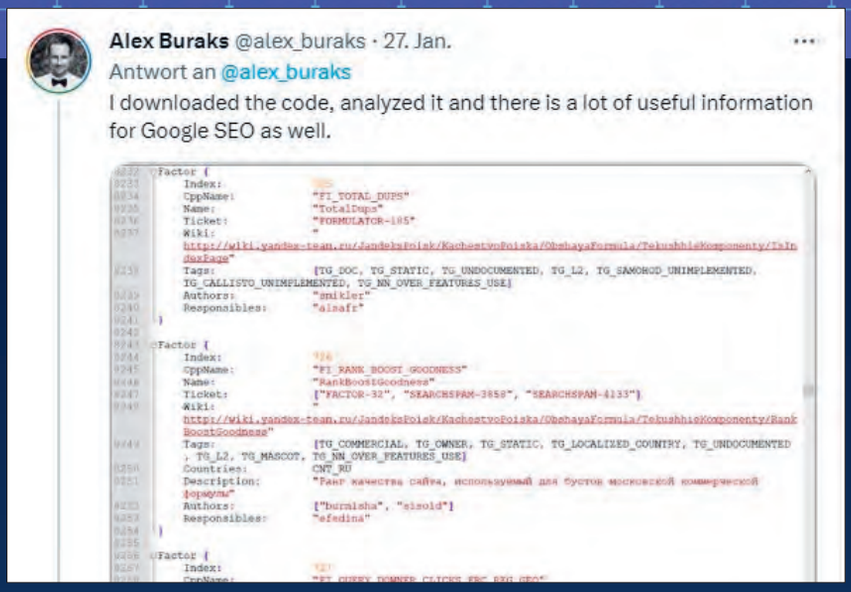

Abb. 1: Analysten sind von den geleakten Daten begeistert *(Quelle: Twitter)*

Auch wenn die größten Suchmaschinen weltweit gesehen unbestritten von Google und Bing betrieben werden, ist die russische Suchmaschine Yandex (der Name entstand dem Vernehmen nach aus „yet another indexer“) auf dem eigenen Markt mit etwa zwei Dritteln Anteil führend. Laut Statista erhält Yandex in Summe mehr als doppelt so viel Traffic wie die chinesische Suchmaschine Baidu. Diese Zahlen schwanken je nach Quelle teilweise stark und am Ende bleibt auch fraglich, inwieweit man den eigenen Angaben aus autokratischen Ländern überhaupt vertrauen mag. Das ist hier aber nicht der Punkt. Es geht darum, wie man mit moderner Technik aus dem Web für Suchanfragen wenigstens die guten Dokumente abfischen kann. Niemand kann sagen oder gar bestätigen, dass die über Yandex geleakten Ranking-Faktoren Rückschlüsse auf Google erlauben bzw. in welchem Ausmaß. Nach dem Bekanntwerden des Lecks beeilte sich Yandex, zu erklären, dass es sich nur um „Fragmente“ aus den Repositorys handeln würde und sich die wirklich verwendeten Signale davon unterscheiden würden. Zur Erklärung sei erwähnt, dass das gesamte Datenleck über 44 GB Daten enthielt. Dass es sich hierbei nur um veraltete interne Daten handeln soll, die keinerlei Anwendung mehr finden, wirkt eher unwahrscheinlich. Allerdings kann es ebenso sein, dass hier tatsächlich nicht alle Faktoren öffentlich vorliegen. Aus den Daten geht allerdings hervor, dass ein Großteil der Faktoren, knapp 990, bereits mit einem Flag versehen sind, das sie als gelöscht oder veraltet kennzeichnet. Somit verbleiben in Summe fast 700 wohl noch aktive Signale.

Die geleakten Daten waren zum Zeitpunkt des Redaktionsschlusses noch immer online abrufbar unter *einfach.st/yandexleak*. Martin MacDonald bietet auf *einfach.st/macdonald* eine Übersicht über diverse Listen, u. a. eine ins Englische übersetzte.

Generell gibt es drei Typen von Ranking-Faktoren. Statische, dynamische und fragespezifische, also solche, die von der eingetippten Suchphrase abhängen und weniger von den Websites selbst.

315 der noch aktuellen Ranking-Faktoren nutzen sogenannte Schwellenwerte (thresholds) für eine SEO-„Überoptimierung“. Überschreitet man einen definierten Grenzwert, reißt man die Latte. Die meisten Signale werden miteinander verrechnet und Scorewerte für Teilbereiche, die ihrerseits aufsummiert werden, bestimmen das Ranking. Bei Grenzwerten gibt es keinen Toleranzbereich und ein Ranking unterbleibt bzw. es gibt eine sogenannte Flag-Strafe. Bekannt sind bei Google derartige Techniken zum Beispiel beim Keyword-Stuffing, also dem allzu häufigen Verwenden eines Keywords im Text. 39 der 315 Signale sind sogenannte „Anti-SEO upper bounds“, das bedeutet, sie verhindern das Ranking aktiv. Solche „Strafen“ greifen für einzelne Keywords oder URLs. Das bedeutet, dass eine Seite für ein bestimmtes Keyword eine Art Rankingsperre haben kann, dieselbe Seite jedoch durchaus für ein anderes Keyword sogar am besten ranken kann. Selbstverständlich kann sich so ein Negativsignal auch auf eine gesamte Domain auswirken.

Bedient sich Yandex bei Google & Co.?

Bereits vor zehn Jahren hielten sich Gerüchte, dass Bing mit bestimmten Keywords die Google-Ergebnisse abfragt, um an aktuelle bzw. neue Dokumente zu kommen, weil der eigene Bot nicht leistungsfähig genug wäre, alles im gleichen Umfang wie Google ständig abzufragen und Neues zu entdecken. Offiziell bestätigt wurde

TIPP

Unter *ipullrank.com (einfach.st/yandex22)* finden Sie ein Formular, über das Sie einen Downloadlink zu einer kompletten Excel-Datei mit allen Ranking-Faktoren bereits übersetzt ins Englische und mit allen Gewichtungsfaktoren anfordern können.

Als .txt-File finden Sie unter *einfach.st/yandex23* bei *webmarketingschool.com* alle Daten flat und direkt im Browser zum Abspeichern.

Noch tiefer gehende Informationen zur Yandex-Architektur gibt es direkt bei Yandex selbst unter *einfach.st/yandex24*.

Wer komplett alle 44,7 GB an Rohdaten haben möchte, findet diese auf Github unter *einfach.st/yandexgithub*.

	A	B	C	D	E
1	CppName	Description(EN)	Rank Coeffici	AntiS	Tags
2	FI_URL_DOMAIN_FRACTION	Coating domain three -bougu and request. (Chelyabinsk lottery - Chelloto. We trans	0,564095297		TG_HOST, TG_DYNAMIC, TG_URL_TEXT, T
3	FI_QUERY_DOWNER_CLICKS_COMBO	factor cunningly combined from FRC and Pseudo-CTR	0,369078039		TG_DYNAMIC, TG_DOWNER, TG_LOCALIZE
4	FI_MAX_WORD_HOST_CLICKS	Domattr clickness for the most expressed word. For example, for all requests in wh	0,345115884		TG_DYNAMIC, TG_DOWNER, TG_USER, TG
5	FI_MAX_WORD_HOST_YABAR	The most characteristic word of the request corresponding to the site, according to	0,315439457		TG_DYNAMIC, TG_DOWNER, TG_USER, TG
6	FI_IS_COM	Domna in Zone .com	0,276250497		TG_HOST, TG_STATIC, TG_URL_TEXT, TG_
7	FI_IS_COM	Domna in Zone .com	0,276250497		TG_NOT_01, TG_HOST, TG_STATIC, TG_UF
8	FI_OQ_BCLM_PLAIN	BCLM on the request of the owners index	0,254915496		TG_DYNAMIC, TG_DOWNER, TG_USER, TG
9	FI_OWNER_CLICKS_PCTR	The owner's clickness regardless of the request	0,231000482		TG_STATIC, TG_OWNER, TG_USER, TG_US
10	FI_OWNER_CLICKS_PCTR	The owner's clickness regardless of the request	0,231000482		TG_STATIC, TG_OWNER, TG_USER, TG_US
11	FI_OWNER_CLICKS_PCTR	The owner's clickness regardless of the request	0,231000482		TG_NOT_01, TG_STATIC, TG_OWNER, TG_
12	FI_MAX_WORD_HOST_RANK	Host Rank according to the most pronounced word of request (usually this is the n	0,230257145		TG_DYNAMIC, TG_DOWNER, TG_LINK_TEX
13	FI_QUERY_DOWNER_CLICKS_PCTR	How often they click in the URLs of this Domainid for this request - Ctr Domainid bl	0,219595036		TG_DYNAMIC, TG_DOWNER, TG_LOCALIZE
14	FI_QUERY_DOWNER_CLICKS_FRC	the ratio of the number of clicks on this Domainid to all clicks on request	0,214713694		TG_DYNAMIC, TG_DOWNER, TG_LOCALIZE
15	FI_QUERY_DOWNER_CLICKS_FRC	SingleOrgMx_38: QueryDOwnerClicksFRC	0,214713694		TG_GEO
16	FI_DOM_PHRASE_CLICK_RANK_BI	Clicking domain on biograms (excluding thesaurus extensions of requests)	0,209866937		TG_DYNAMIC, TG_DOWNER, TG_USER, TG
17	FI_OWNER_REQS_POPULARITY	The popularity of Owner is in requests	0,209508534		TG_STATIC, TG_OWNER, TG_USER, TG_US
18	FI_OWNER_REQS_POPULARITY	The popularity of Owner is in requests	0,209508534		TG_STATIC, TG_OWNER, TG_USER, TG_US
19	FI_OWNER_REQS_POPULARITY	The popularity of Owner \ \'And in the requests	0,209508534		TG_NOT_01, TG_STATIC, TG_OWNER, TG_
20	FI_HAS_NO_QUERY_SHOWS	the request has no shows	0,205699196		TG_QUERY_ONLY, TG_BINARY, TG_OFTEN
21	FI_HAS_NO_QUERY_SHOWS	For this request, there is no information about the clickness of 1 - there is no reque	0,205699196		TG_DYNAMIC, TG_QUERY_ONLY, TG_LOCA
22	FI_DOM_PHRASE_YABAR_BI	Transitions to the site from search engines by biograms, according to the bar (excl	0,205184905		TG_DYNAMIC, TG_DOWNER, TG_USER, TG
23	FI_QUERY_DOWNER_WEIGHT_CLICK	w/k	0,202186194		TG_DYNAMIC, TG_DOWNER, TG_LOCALIZE
24	FI_OWNER_NAV_QUOTA	The share of clicks for navigation requests	0,18974311		TG_STATIC, TG_OWNER, TG_USER, TG_US
25	FI_OWNER_NAV_QUOTA	The share of clicks for navigation requests	0,18974311		TG_STATIC, TG_OWNER, TG_USER, TG_US
26	FI_OWNER_NAV_QUOTA	The share of clicks for navigation requests	0,18974311		TG_NOT_01, TG_STATIC, TG_OWNER, TG_
27	FI_QUERY_DOWNER_ONLY_CLICK_RA	o/i	0,185032224		TG_DYNAMIC, TG_DOWNER, TG_LOCALIZE
28	FI_PAGE_RANK	Page Rank. The factor will be remarked.	0,182867833	1	TG_DOC, TG_LINK_GRAPH, TG_STATIC, TG
29	FI_QUERY_DOWNER_ONLY_CLICK_RA	o/i	0,179216994		TG_DYNAMIC, TG_DOWNER, TG_LOCALIZE
30	FI_SUBQUERY_THEME_MATCH_A	Coincidence of thematic spectra of request and document. Request themes - the r	0,178646516		TG_DOC, TG_DYNAMIC, TG_THEME_CLASS
31	FI_OWNER_CLICKS_PCTR_REG	The owner's clickness regardless of the request, separately in the regions	0,166327421		TG_STATIC, TG_OWNER, TG_LOCALIZED_
32	FI_OWNER_CLICKS_PCTR_REG	The owner's clickness regardless of the request, separately in the regions	0,166327421		TG_NOT_01, TG_STATIC, TG_OWNER, TG_
33	FI_HAS_DETERMINED_CITIES	The city is defined for the site	0,165031404		TG_DOC, TG_STATIC, TG_STATIC_REGINF
34	FI_QUERY_DOWNER_CLICKS_COMBO	Query Download Clicks Combo, in small regions from Relev_regions.web.txt	0,160420714		TG_DYNAMIC, TG_DOWNER, TG_LOCALIZE
35	FI_HAS_NO_QUERY_DOWNER_SHOWS	For this Domainid for this request, there is no information about clickability 1 - requ	0,160379345		TG_DYNAMIC, TG_DOWNER, TG_LOCALIZE
36	FI_REG_HOST_RANK	It reads in the same way as the Hostrank factor, but not on all the Owner graph, but	0,15671244		TG_LOCALIZED_COUNTRY, TG_LINK_GRA
37	FI_REG_HOST_RANK	It reads in the same way as the Hostrank factor, but not on all the Owner graph, but	0,15671244		TG_NOT_01, TG_LOCALIZED_COUNTRY, T
38	FI_QUERY_DOWNER_WS_MAX_WHR_A	The amount of factors 192 and 341 with scales 0.298942 and 0.454625, respective	0,152953809		TG_DEPRECATED, TG_DYNAMIC, TG_DOW
39	FI_QUERY_DOWNER_SATISFIED4_RATI	r_s4b/(r_k + 10)	0,148292223		TG_DYNAMIC, TG_DOWNER, TG_LOCALIZE
40	FI_QUERY_DOWNER_YABAR_VISITS		0,147136648		TG_DYNAMIC, TG_DOWNER, TG_USER, TG
41	FI_OQ_BM25_LEM	BM25 on the request for Domattr index	0,12966893		TG_DYNAMIC, TG_DOWNER, TG_USER, TG
42	FI_FIRST_WORD_HOST_CLICKS	The clickability of the host according to the first word of the request. Quite often, the	0,129641402		TG_DYNAMIC, TG_DOWNER, TG_USER, TG
43	FI_YABAR_HOST_AVG_ACTIONS	The average for users is the number of active actions (clicks, clicks) with the contin	0,12797973		TG_STATIC, TG_HOST, TG_USER, TG_BRO
44	FI_YABAR_HOST_AVG_ACTIONS	The average for users is the number of active actions (clicks, clicks) with the contin	0,12797973		TG_STATIC, TG_HOST, TG_USER, TG_BRO
45	FI_YABAR_HOST_AVG_CTIONS	The average for users is the number of acti e actions (cli ks, clicks) with the contin	0,12797973		TG_NOT_01, TG_STATIC, TG_HOST, TG_US
	FI_OWN R_SESS DUR	ID/K normalized time to click	16		TG_STATI TG WNER, TG_USER, TG_US
	FI_ R_SESS	D/K normalized time to lick	E		STAT T IER, TG_USER, TG_U
	FI_ R_SESS	D/K normalized time t K			NOT IC, TG W ER, TG
	25_ST	I2F request ndex			VN VNEF ER,
	_DOV	nour			IF P

Abb. 2: Die komplette Liste aller Ranking-Faktoren gibt es im Web im Excel-Format.

das allerdings nie. Nun scheint es so, als würde wohl auch Yandex bei Google scrapen, dessen Tensor Flow, BERT und MapReduce als Technologie nutzen sowie diverse Linkmetriken. Den Daten zufolge scrapt Yandex aber auch Bing, YouTube und TikTok.

Hat der „PageRank" einen starken Einfluss?

Für viele war es wohl eine echte Überraschung, dass Yandex offenbar den PageRank-Algorithmus benutzt. Zur Erinnerung: Dies war das Herzstück beim Markteintritt von Google. Erstmals hat damals eine Suchmaschine nicht nur den Inhalt einer Seite bewertet, sondern eben auch die Links von außen (Backlinks) bzw. von anderen Domains dorthin. Die Idee dahinter zu einer Zeit, in der Links von Webmastern noch per Hand gesetzt bzw. programmiert wurden, war, dass gute Inhalte mehr Hinweise via Link bekom-

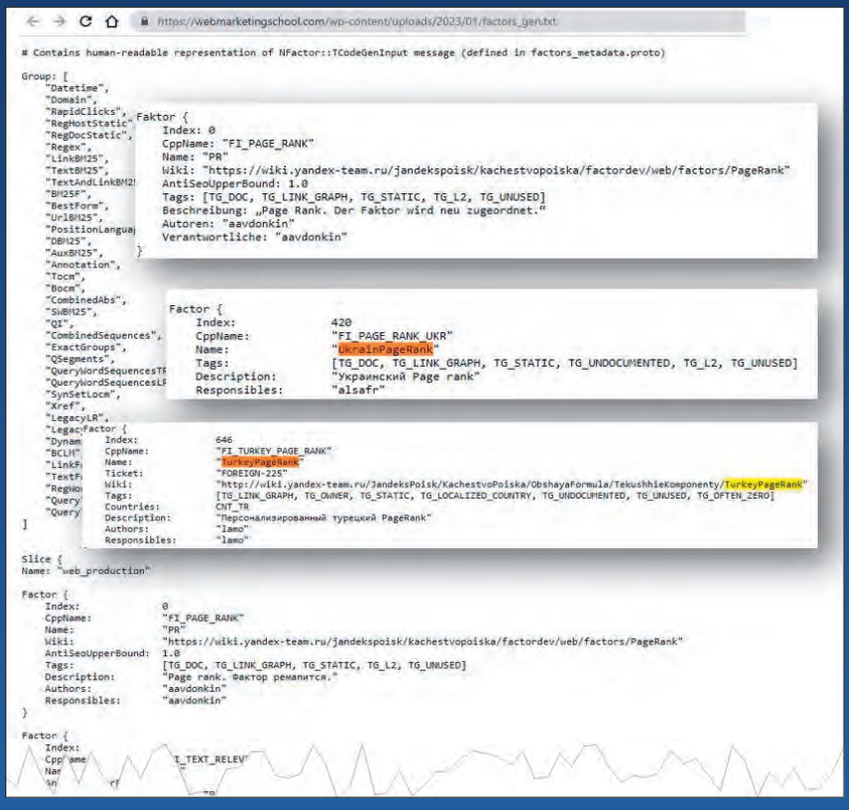

Abb. 3: Für die Türkei und die Ukraine wird offenbar ein gesonderter PageRank berechnet *(Quelle: Yandex-Datenleak, einfach.st/yandexleak).*

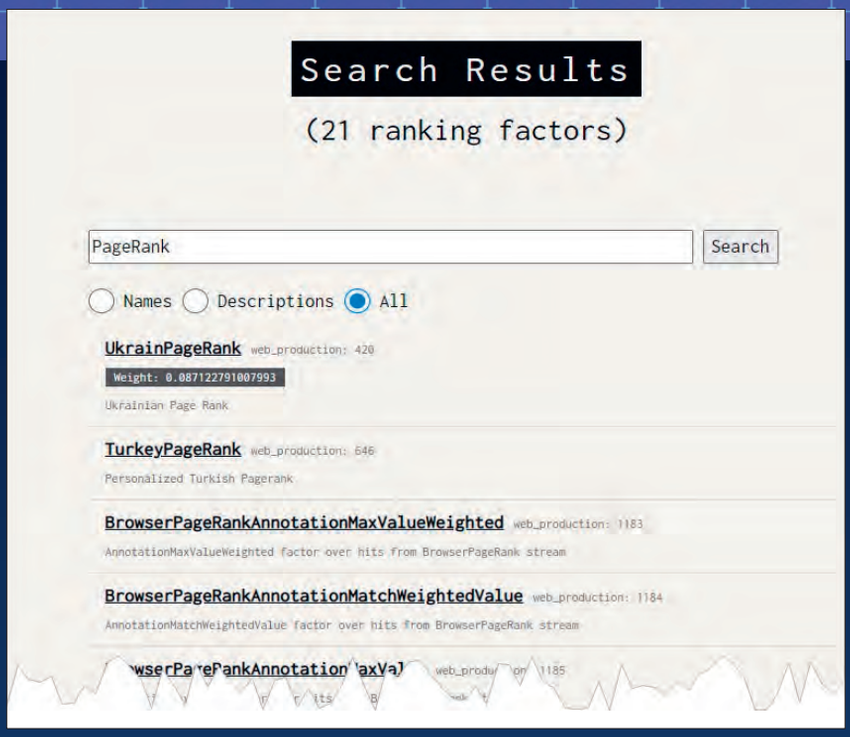

Abb. 4: Unter yandex-explorer.herokuapp.com lässt sich nach Ranking-Faktoren suchen (hier nach PageRank).

men als unnütze. Zwei Doktoranden der Stanford University wollten wissen, wer im Web auf ihre Forschungsergebnisse verlinkt. Da es keine zentrale Instanz dafür gab und bis heute nicht gibt, kamen sie auf die Idee, das gesamte Internet (damals ca. 25 Millionen Seiten) zu downloaden und manuell nach den Verweisen zu ihnen zu suchen. Dabei stellten sie durch Zufall fest, dass die Sites, die sich selbst häufig im Web benutzten, auch die meisten Backlinks hatten. Schnell war die Idee geboren, diesen Faktor in das Ranking einer Suchmaschine einzubeziehen. Nur leider wollte niemand diese Idee verwenden und so waren beide „gezwungen", selbst eine Suchmaschine zu bauen. Sergey Brin und Larry Page warfen also ihr Doktorandenstudium hin und der Rest ist Geschichte. Die Idee des PageRanks wurde allerdings vorher zum Patent angemeldet und gehört seither der Stanford University. Google zahlt dem Vernehmen nach jedes Jahr einen dreistelligen Millionenbetrag an die Uni für die Nutzung. Insofern dürfte Yandex das patentrechtlich geschützte Verfahren möglicherweise gar nicht einfach so nutzen, es sei denn, Stanford würde hierfür eine Lizenz an den russischen Betreiber vergeben. Dieser Faktor bleibt bisher im Web unbeachtet und man darf gespannt sein, ob dies juristische Konsequenzen haben wird – sofern man aktuell einem russischen Unternehmen juristisch überhaupt beikommen kann.

Yandex geht sogar noch einen Schritt weiter und berechnet offenbar für einzelne Länder einen eigenen PageRank, zumindest für die Türkei und die Ukraine. Schon vor vielen Jahren wurde übrigens hinter vorgehaltener Hand kolportiert, dass man für den Großraum in und um Moskau die Backlinks für das Ranking außen vor lassen würde. Die dort ansässigen Linkspammer waren wohl so aktiv, dass die Suchergebnisse für diese Gegend qualitativ besser waren, wenn man die Backlinks nicht einbezogen hat.

89 der geleakten Signale haben wohl die Aufgabe, Spam bei Links zu erkennen. Das erscheint als durchaus sehr differenziert und zeigt erneut, Link ist nicht gleich Link und es wird heute immer schwerer, den wahren Wert eines Links nachzuvollziehen.

Einige positive Faktoren

FI_URL_DOMAIN_FRACTION (Gewichtung: +0,5640952971)

Yandex misst mit diesem Wert, wie stark eine Suchphrase mit der Domain der URL übereinstimmt. Um diesen Wert zu berechnen, nutzt man anscheinend je drei aufeinanderfolgende Buchstaben und ermittelt, welcher Anteil aller je drei Buchstabenkombinationen im Domainnamen enthalten ist.

FI_QUERY_DOWNER_CLICKS_COMBO (Gewichtung: +0,3690780393)

Die Bedeutung dieses Faktors ist noch unklar. Übersetzt bedeutet er in etwa „schlau kombiniert aus FRC und Pseudo-CTR". Was mit der Abkürzung „FRC" gemeint ist, bleibt offen.

FI_MAX_WORD_HOST_CLICKS (Gewichtung: +0,3451158835)

Dieser Faktor berücksichtigt für das wichtigste Wort einer Domain, inwieweit es in der Suchphrase enthalten ist und wie oft im positiven Fall genau auch auf diese Domain geklickt wird. Ein Beispiel: Wie oft wird bei Suchanfragen, die „Website Boosting" enthalten, auf das Ergebnis für eine Domain geklickt, das als Hauptbegriff auch „Website Boosting" enthält?

FI_MAX_WORD_HOST_YABAR (Gewichtung: +0,3154394573)

Übersetzt steht im Kommentar dazu in etwa „das charakteristischste Suchwort, das der Website entspricht, laut Bar." Offenbar spielt es hier eine Rolle, welches Suchwort am meisten über die Yandex-Toolbar-Suche mit einer Website verbunden ist.

FI_REG_HOST_RANK (Gewichtung +0,1567124399)

Yandex nutzt hier eine Art übergreifenden Domain-Ranking-Faktor. Einzelne URLs einer in Summe gut rankenden Domain bekommen also einen Bonus. Eigentlich begünstigt sich dieser Faktor im Lauf der Zeit selbst bzw. verstärkt den „Domaineffekt" immer weiter. Ob und wie dieser Effekt gedämpft wird, darüber findet man nichts Genaues in den Daten.

FI_IS_COM (Gewichtung: +0,2762504972)

Dieser Faktor sorgt tatsächlich dafür, dass .com-Domains im Ranking bevorzugt werden!

FI_IS_NOT_RU (Gewichtung: +0,08128946612)

Der Faktor zeigt, dass es als gut bewertet wird, wenn die Domain keine .ru ist!. Anscheinend vertraut die russische Suchmaschine ausgerechnet Seiten aus dem eigenen Land nicht besonders.

Einige stark negative Faktoren

FI_ADV (Gewichtung: –0,2509284637)

Dieser Faktor bestimmt, ob Werbung jeglicher Art auf der Seite vorhanden ist, und verhängt dafür die am stärksten gewichtete Strafe für einen einzelnen Ranking-Faktor. Allerdings wird Werbung von Yandex bzw. solche aus dem Yandex-Netzwerk (ähnlich Google Ads) offenbar anders und wohlwollender behandelt.

FI_DATER_AGE (Gewichtung: –0,2074373667)

Dies ist die Differenz zwischen dem aktuellen Datum und dem durch eine Datumsfunktion ermittelten Datum des Dokuments. Der Wert ist eins, wenn das Dokumentdatum von heute ist, null, wenn das Dokument zehn Jahre oder älter ist oder wenn das Datum nicht definiert/ermittelbar ist.

FI_QURL_STAT_POWER (Gewichtung: –0,1943768768)

Hier geht es um die Anzahl der URL-Impressionen in Bezug auf eine Suchabfrage. Wenn eine URL in vielen Suchen auftaucht, wird sie bestraft bzw. bekommt einen Abzug für das Ranking – das fördert die Vielfalt der Ergebnisse.

FI_COMM_LINKS_SEO_HOSTS (Gewichtung: –0,1809636391)

Der Faktor berücksichtigt den Prozentsatz der eingehenden Links mit „kommerziellem" Ankertext. Der Faktor wird auf 0,1 vermindert, wenn der Anteil solcher Links > 50 % beträgt, ansonsten wird er auf null gesetzt. Welche Ankertexte kommerziell bzw. werthaltig sind, kann Yandex genauso wie Google aus dem eigenen Werbesystem entnehmen. Je mehr Werbetreibende für einen Klick bezahlen, desto werthaltiger ist dieses Keyword.

Was ist mit den Klickdaten auf Suchergebnisse?

Yandex nutzt wohl in besonderem Maße und seit Langem Verhaltensdaten (CTR, Zeit auf der Seite, Absprungrate, Klicks, die allgemeine Seitenaktivität, sogar die Anzahl, wie viele User sich ein Bookmark von einer Seite setzen, und einiges mehr). Verschiedene Auswertungen der Ranking-Faktoren zeigen sogar an, dass die User-Metriken in Summe zumindest prozentual den höchsten Anteil am Ranking haben.

Google hat das bisher immer dementiert. Dies ließe sich sicher auch schlecht erklären, weil sofort „Überwachung" und „Datenkrake" als Überschriften zu lesen wären. Aber machen wir uns nichts vor, für das Ranking sind gerade diese Werte enorm wichtig, drückt sich doch in ihnen die User-Zufriedenheit aus.

Google hat das immer wieder öffentlich abgestritten. Und dies, obwohl es zwar geschwärzte, aber öffentlich einsehbare Dokumente wie den FTC Report vom 8. August 2012 gibt, in dem es zum Beispiel auf Seite 14 heißt: *„The Ranking itself ist affected by the click data, If we discover that, for a particular query, hypothetically, 80 percent of people klcik on Result No. 2 an only 10 percent click on Result No. 1, after a while we figure out, well, rpobobaly Result 2 is the one people want. So we'll switch it."* Auf Seite 16 steht weiter: *„Google tracks user click-through rates (and relies on such click-through data to improve its web search results in a number of ways."* Und auf S. 82 dann: *„based on the crawled text, the pages are rated using such factors as click-through rates."* Und das US-Patent 9.727.653 B2 eingereicht von Google und datiert auf den 8. August 2017 trägt den nachdenklichen Namen *„System and method for indentifiying and ranking user preferences"*.

Letztlich ist aber auch das natürlich ehrlich gesagt noch immer kein endgültiger Beweis für die Verwendung von User-Daten für das Ranking. Bei Yandex gab es schon mehrere positive Versuche, über Klickmanipulationen Ergebnisse nach vorne zu bringen. Insofern sind die diesbezüglichen Daten nur noch die letzte Bestätigung, dass die Klickrate in den Suchergebnissen dort Einfluss hat. Dass solche Manipulationen bei Google nur sehr viel schwerer zu machen sind, ist klar. Google hat über den Browser Chrome mit über 60 % Marktanteil und eingeloggten Usern (mit deren Klickhistorie) ganz andere Möglichkeiten, „ungewöhnliche" Klickmuster zu erkennen und deren Einfluss bei einem algorithmisch vermuteten Missbrauch zu unterbinden. Yandex versucht dem Vernehmen nach, allerdings auch

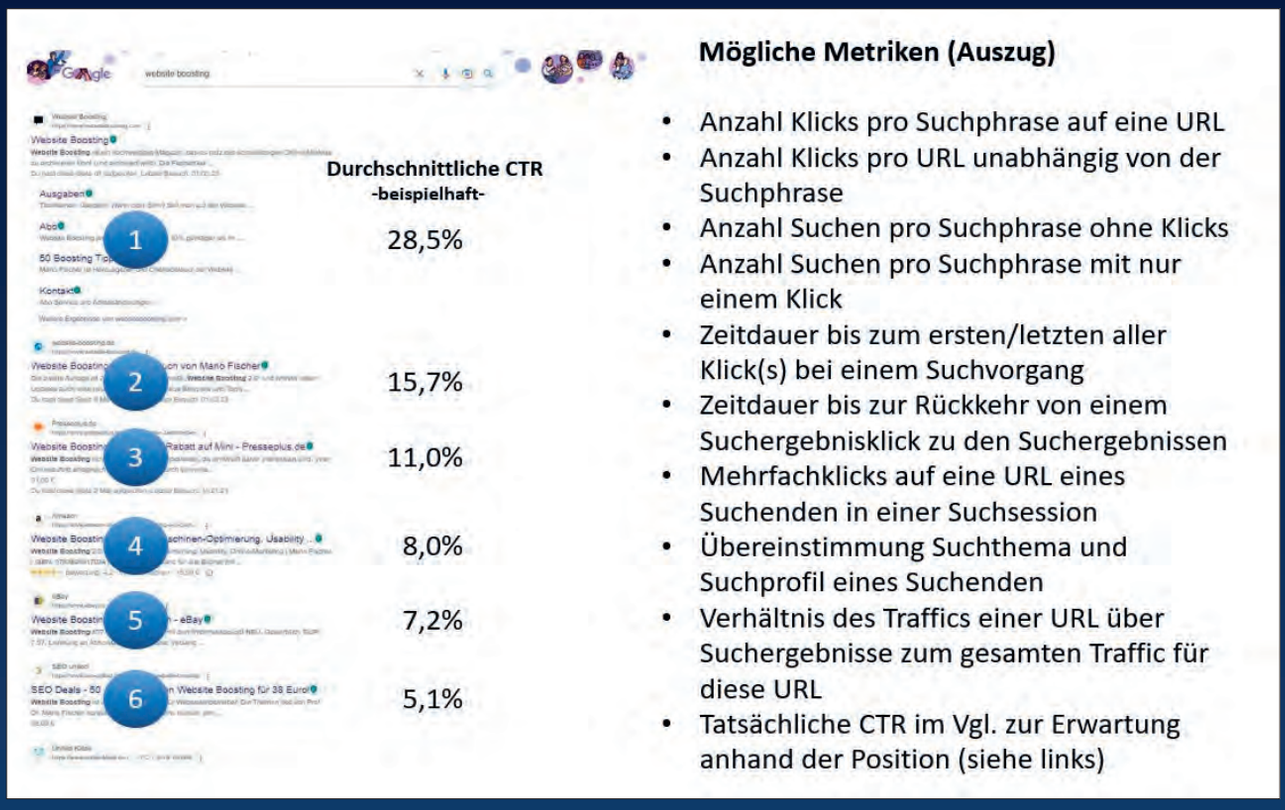

Abb. 5: Welche Metriken könnten in einem Suchergebnis verwendet werden?

maschinell zu erkennen, ob Klicks auf Ergebnisse in Suchergebnissen künstlich manipuliert sind, und verhängt empfindliche Strafen oder auch einen vollständigen Ausschluss aus dem Ranking.

Abbildung 5 zeigt beispielhaft, welche Metriken Suchmaschinen bei der Nutzung ihrer Suchergebnisse erfassen können bzw. Yandex laut dem Leak tatsächlich erfasst und auswertet. An folgendem Beispiel wird die Nützlichkeit eines solchen Vorgehens mehr als deutlich. Die Klickraten (CTR) sind natürlich bei besseren Positionen deutlich höher und nehmen insofern natürlich deutlich ab, je weiter unten ein Ergebnis auftaucht. Dies muss man bei Auswertungen des User-Verhaltens selbstverständlich berücksichtigen. Die Daten aus dem Beispiel in der Abbildung stammen von einer Analyse des Toolanbieters SISTRIX aus dem Jahr 2020. Angenommen für eine Suchphrase wird Platz drei deutlich häufiger angeklickt, als es aufgrund der Position zu erwarten wäre. Die tatsächlich CTR liegt z. B. bei 45 % – statt bei den zu erwartenden 11 %. Die Gründe für diese „Bevorzugung“ durch Suchende können vielfältiger Natur sein. Platz eins und zwei passen laut Title und Description oder wegen der Domain/URL vielleicht nicht wirklich oder man hat früher schon einmal dorthin geklickt und nicht gefunden, was man gesucht hat. Was auch immer der Auslöser für diese Anomalie ist, die Suchenden mögen Ergebnis drei deutlich lieber. Käme jetzt noch ein längerer Aufenthalt des Suchenden auf der Zielseite dazu (sogenannter Long Click) bzw. kehren diese nicht sofort wieder zurück (sogenannter Short Click) und klicken auf ein anderes Ergebnis, erscheint es relativ wahrscheinlich, dass Ergebnis drei am Ende dann doch ein besserer Treffer (laut „User-Votum“) für eine Suchphrase ist als der vormals (laut Algorithmen) errechnete Platz eins und zwei. Würde eine aufmerksame Suchmaschine dies einfach ignorieren und das offenbar mehr gewollte Ergebnis von Platz drei dort stehen lassen oder würde man es nach oben heben? Natürlich dürfte eine solche Ranking-Veränderung nicht schon nach wenigen Klicks greifen, sondern braucht statistische Relevanz und sicher auch anschließende A/B-Tests zur Verifizierung. Wie erwähnt: Google verneint dies öffentlich, spricht von „noisy signals“ (also von zu viel Störsignalen) und sagt damit eigentlich, dass es egal wäre, wie gut die Suchenden die Ergebnisse finden, die Ranking-Algorithmen wissen es besser. Das ist in der Tat für einen Betrachter von außen schwer zu glauben oder zu verstehen. Yandex fährt offenbar sehr gut mit der Nutzung dieser Usersignale.

Zusammenfassende Erkenntnisse

Szymon Stowik hat sich intensiv mit dem Datenleak von Yandex beschäftigt und bisher herausgefun-

den, dass dort wohl einer der wichtigsten Ranking-Faktoren der Page Rank (länderspezifisch) ist. Ebenso spielt das Alter der Links eine Rolle. Am Rande sei erwähnt, dass es auch hierzu bereits seit 2011 ein Patent von Google gibt (US-Patent 8549014 B2). Weitere Ranking-Faktoren sind der Traffic sowie der Prozentsatz des organischen Traffics am gesamten Traffic. Zahlen in URLs werden negativ bewertet, ebenso wie viele Schrägstriche in URLs. Das bedeutet nichts anderes als die Verzeichnistiefe, also je tiefer, desto schlechter. Yandex legt wohl großen Wert auf die Host-Stabilität und die Hygiene: Je weniger 5xx- und 4xx-Fehler, desto besser für das Ranking.

Das Alter des Dokuments (steht ebenso im Google-Patent US-Patent 8549014 B2) und die letzte Aktualisierung sind wichtig für eine gute Bewertung. Yandex geht offenbar so weit, sogar die CTR für ähnliche Suchanfragen (Synonyme etc.) zu berücksichtigen. Ebenso findet die Anzahl der Aufrufe einer bestimmten URL pro Suchanfrage Eingang in das Ranking. Keinerlei Überraschung ist es allerdings, dass die thematische Relevanz – ebenso wie bei Google – besonders wichtig ist. NLP-Analysen werden im Kontext der Natürlichkeit des Textes für die russische Sprache verwendet. Der Mechanismus soll Inhalte erkennen, die von einem Synonymisierer oder einem Automaten generiert worden sein könnten. Mit anderen Worten versucht Yandex, zu ermitteln, wie unnatürlich der Text aus Sicht der russischen Sprache ist. Analysiert werden verschiedene Inhaltselemente wie Wortlänge, Anzahl der Verben, Pronomen und andere Wortarten. Die Länge des Textes wirkt sich ebenfalls positiv auf das Ranking aus. Die verwendeten Wörter müssen als Voraussetzung selbstverständlich semantisch und nach bestimmten Grammatikregeln sinnhaft sein und nicht etwa einen zusammengewürfelten Wortsalat darstellen.

Enthält die URL eine Länderbezeichnung oder einen Stadtnamen, wird dies entsprechend berücksichtigt, sofern die Suchabfrage Gleiches enthält und/oder die Suchenden sich in diesen Gebieten aufhalten. Dass „sprechende" URLs gut für SEO sind, ist mittlerweile schon lange kein Geheimnis mehr. Hier bestätigt sich die allgemeine Vermutung nochmals in Form von Geo-Faktoren. Yandex hat eine ganze Menge an URL-bezogenen Ranking-Faktoren, was die Empfehlung nochmals verstärkt, die URL-Struktur(en) wohlbedacht zu wählen und nicht einem CMS oder Shopsystem nach „Product-IDs" oder anderen, nichtssagenden Zählern zu überlassen.

Yandex versucht auch, den Seitentyp wie zum Beispiel Shop, Blog, News, Vergleichsseite etc. zu ermitteln und zu verwenden. Ebenso findet Technik wie etwa das verwendete CMS Eingang in die Analyse.

Sind bei einer Abfrage nicht alle Wörter auch im Dokument enthalten, rankt dieses umso schlechter. Umgekehrt: Je mehr Wörter sich im Dokument finden, desto besser. Dies kennt man ebenfalls von Google. Dort ist die Nähe von Wörtern bei sogenann-

INFO

Was bedeutet DSSM (Deep Semantic Similarity Model)?

Die semantische Ähnlichkeit ist eine Metrik, die über eine Reihe von Dokumenten oder Begriffen greift, wobei die Idee der Distanz zwischen ihnen auf der Ähnlichkeit ihrer Bedeutung oder ihres semantischen Inhalts basiert. Ganz im Gegensatz zu der Art Ähnlichkeit, die hinsichtlich ihrer syntaktischen Darstellung (zum Beispiel ihres Zeichenfolgenformats) vorhergesagt werden kann. Das Wort „Zug" für Eisenbahn und bei Soldaten ist syntaktisch gleich, aber semantisch verschieden. „ICE" und „Eisenbahn" sind dagegen syntaktisch verschieden, bedeuten semantisch jedoch (fast) das Gleiche.

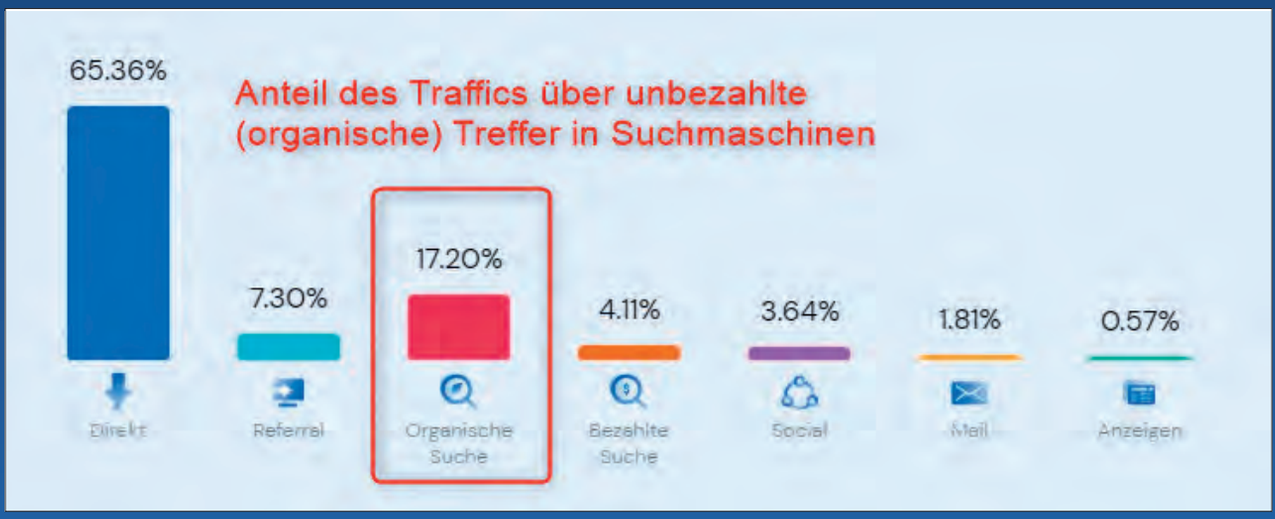

Abb. 6: Vermutlicher Ranking-Faktor bei Yandex: Der Traffic-Anteil über Suchmaschinen (hier amazon.de, gemessen von similarweb.com)

ten Long-Tail-Abfragen ein wesentlicher Relevanzfaktor (US-Patente 20020143758 A1 und 20080313202 A1). Yandex wertet die Übereinstimmung prozentual aus.

Anhand von Who-is-Einträgen wird die Wahrscheinlichkeit bewertet, dass das Hosting zu einem Spammer gehört. Ebenso wird nach einer Teilnahme an Linknetzwerken gesucht und die Sites/Seiten werden entsprechend aussortiert. Einige Faktoren zielen darauf ab, die Anzahl schlechter Links zwischen zwei Hosts zu ermitteln und negativ für das Ranking einfließen zu lassen.

Es ist spannend, zu sehen, dass auch Yandex DSSM (siehe Kasten) verwendet. Dieses Modell wird in Googles Tensor Flow abgebildet, das Yandex dem Leak gemäß offenbar ebenfalls verwendet. Unter anderem wird die DSSM-Wahrscheinlichkeitsvorhersage anhand der Dokument-URL und des Titels genutzt, um zu ermitteln, wie viele Produkte es auf der Seite gibt. Das ergibt durchaus Sinn, insbesondere wenn festgestellt wird, dass mehrere Produkte zum Beispiel auf einer typischen Kategorieseite in Online-Shops besser geeignet sind, um einen besseren Preis-Leistungs-Überblick zu bekommen. Eine einzelne Produktseite bietet diesen Vorteil nicht. Je nach Formulierung einer Suchanfrage „Welche Produkte … kaufen“ vs. „Produkt XY kaufen“ kann dann der richtige Seitentyp ranken. Tatsächlich wird DSSM in 135 Ranglistenfaktoren verwendet, sodass man von einem durchaus dominanten Faktor ausgehen kann. Auch der bekannte BM25-Algorithmus (BM steht für „best matching“) wird anscheinend zur Textanalyse eingesetzt. Gleich 33 verschiedene Faktoren beinhalten ihn bzw. seine Bezeichnung. Dazu zählen insbesondere auch WDF-IDF (bzw. TF-IDF) ähnelnde Ansätze. Damit wird ermittelt, wie häufig ein bestimmtes Wort in einem Dokument (WDF – within document frequency) und zugleich wie selten es in allen anderen Dokumenten vorkommt (IDF – inverse document frequency). Das bedeutet nichts anderes als die Messung, wie stark ein Wort aus einem Text herausticht und wie gut es sich gleichzeitig als bezeichnendes Suchwort eignet (je seltener woanders, desto besser).

Viele weitere Faktoren zielen auf die Analyse von Text ab. So ist „FI_ADV_PRONOUNS_PORTION“ für die Ermittlung des Anteils an Pronomen auf der Seite zuständig. „FI_PERCENT_FREQ_WORDS“ bewertet den Prozentsatz der Wörter, die zu den 200 häufigsten Wörter im Vergleich zum restlichen Text gehören. Besonders bei „FI_AURA_DOC_LOG_AUTHOR“ geht es darum, zu ermitteln, für wie viele Dokumente der Domaineigentümer auch tatsächlich als Autor erkannt werden kann. Mit einer negativen Gewichtung von –0,1339319854 geht „FI_CLASSIF_IS_SHOP“ in das Ranking ein. Yandex mag ganz offensichtlich keine Shopseiten.

Weiterhin enthält der Leak etwa zehn Faktoren, die darauf hindeuten, dass die Tageszeit und der Wochentag einer Suche einen Einfluss auf das Ranking haben. Das ist allerdings wenig überraschend.

Anders als Google hat Yandex kein eigenes Rendering-System für JavaScript. Somit beschränkt man sich – anders als bei Google – rein auf das textbasierte Crawlen und nicht auf durch CSS und JavaScript induzierte Veränderungen an (sichtbarem) Text und Layout.

Fazit

Die geleakten Ranking-Faktoren von Yandex erscheinen durchaus sinnvoll und wirken praktikabel. Viele davon wurden und werden von Experten auch so oder so ähnlich bei Google vermutet. Man darf getrost davon ausgehen, dass Google sogar noch detaillierter vorgeht und in Summe betrachtet wahrscheinlich sogar noch mehr Faktoren einsetzt. Das ist am Ende alles eine Frage der Rechenpower, des Speicherplatzes und der Wirtschaftlichkeit. Besonders spannend ist wohl, dass und wie viele Faktoren es bei Yandex gibt, die das User-Verhalten aufzeichnen, auswerten und für das Ranking verwenden.

Einige Experten haben übrigens versucht, über Ranking-Vergleiche zwischen Google und Yandex herauszubekommen, ob es ähnliche Treffer gibt. Diese Tests zeigten zwar gewisse Übereinstimmungen, aber die Abweichungen waren dann doch zu groß, um Rückschlüsse auf ähnliche Ranking-Faktoren zu erlauben. Allerdings muss stark bezweifelt werden, ob diese Methode wirklich aussagekräftig ist. Einige Faktoren bei Yandex, wie die Bevorzugung von .com-Domains oder Boni für Sites wie Wikipedia, gibt es bei Google mit an Sicherheit grenzender Wahrscheinlichkeit nicht. Zudem gibt Yandex wohl auch Backlinks noch immer starkes Gewicht, während diese Dominanz bei Google schon seit Längerem spürbar zurückgeht. Dies und einige weitere Gründe können durchaus unterschiedliche Seiten nach oben priorisieren, auch wenn viele andere Ranking-Faktoren durchaus übereinstimmen könnten, wenn auch nicht in gleicher Gewichtung. So bleiben am Ende nur Vermutungen, auch wenn diese durchaus ziemlich glaubwürdig erscheinen. ¶