Ist die Webseite suchmaschinenoptimiert?

Vorstellung eines Online-Tools zur Analyse der Wahrscheinlichkeit der Suchmaschinenoptimierung auf einer Webseite

Sebastian Sünkler, Dirk Lewandowski

Hamburg University of Applied Sciences, Department of Information,
Finkenau 35, 22081 Hamburg (Germany)
{sebastian.suenkler, dirk.lewandowski}@haw-hamburg.de

Abstract

Das SEO-Tool ist eine Webanwendung, die die Wahrscheinlichkeit von Suchmaschinenoptimierung (SEO) auf einer Webseite ermittelt. Für die Berechnung der Wahrscheinlichkeit werden insgesamt 20 Merkmale halb-automatisch erhoben und in drei Prozessen ausgewertet. Dafür analysiert das Tool zuerst den Quelltext der URL auf Informationen über die Verwendung von SEO-Plugins und Analytics Tools. Zweitens bestimmt es die Kategorie der gegebenen URL anhand manuell klassifizierter Websites und schließlich berechnet es verschiedene technische und inhaltliche SEO-Indikatoren. Die Ergebnisse aus diesen Prozessen bilden anschließend die Basis für die Einordnung der URL anhand eines regelbasierten Klassifikators. Die Demo des Tools ist unter http://5.189.155.20:5000/ verfügbar.

Keywords: Softwaredemonstration; Suchmaschinen; Suchmaschinenoptimierung; SEO; Datenanalyse

1 Einleitung

Inhaltsanbieter im Web sind auf Top-Positionen im Ranking in kommerziellen Suchmaschinen wie Google angewiesen, da sie dadurch einen großen Anteil ihres Website-Traffics generieren.

Eine wichtige Möglichkeit, um gute Platzierungen zu erhalten, ist die Suchmaschinenoptimierung (*search engine optimization*, SEO) (Li et al., 2014). Durch die hohe Relevanz von SEO für die Sichtbarkeit der Anbieter, die erheblichen Mittel, die in SEO-Maßnahmen fließen (McCue, 2018), sowie nach Aussagen von Expert/innen (Schultheiß/Lewandowski, 2020) kann davon ausgegangen werden, dass suchmaschinenoptimierte Inhalte einen erheblichen Einfluss auf die Suchergebnisseiten haben. Dieser Einfluss ist bisher kaum erforscht. Mit halb-automatisierten Prozessen und einem regelbasierten Ansatz in der Umsetzung in einem Software-Framework soll dieser Effekt erforscht werden. Ein Teil dieses Frameworks ist ein Online-Tool zur Klassifizierung der Wahrscheinlichkeit von SEO auf einer bestimmten Webseite, um schnelle Tests durchzuführen. Die Demo zu dem Tool ist unter http://5.189.155.20:5000/ verfügbar.

2 Prozesse zur Identifikation von SEO auf einer Webseite

Die Identifikation von SEO auf einer Webseite basiert zum einen auf relevanten Indikatoren für Suchmaschinenoptimierung und auf einer Klassifikation der URL anhand eines Abgleichs mit Websites, die in Kategorien eingeteilt sind, die auf optimierte und nicht optimierte Websites hinweisen. Insgesamt werden 20 Merkmale geprüft. Der Prozessablauf ist in Abbildung 1 dargestellt und kann in drei Schritte unterteilt werden:

1. Abrufen der URL, um den HTML-Code und die Metadaten zu extrahieren,
2. Generierung des Inputs für den regelbasierten Klassifikator auf drei Stufen,
3. Bestimmen der Wahrscheinlichkeit von SEO.

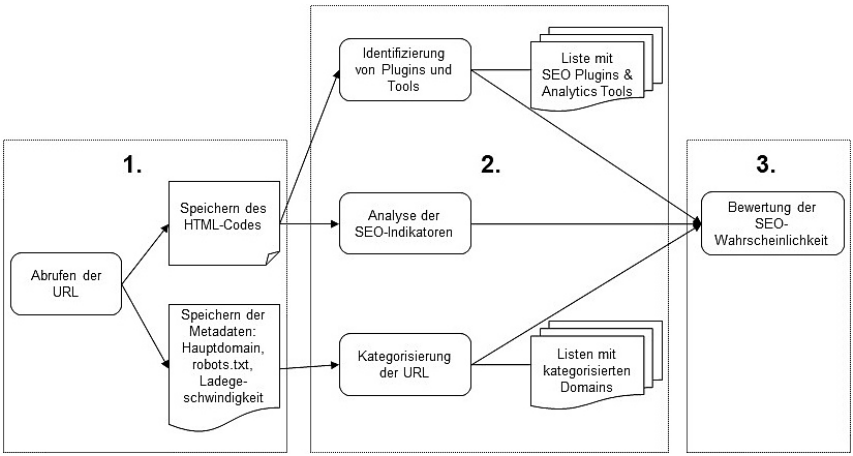

Abb. 1 Prozesse für die Bestimmung von SEO auf einer Webseite

2.1 Abrufen der URL und Speichern der Metadaten und des HTML-Codes

Im ersten Schritt wird eine eingegebene URL heruntergeladen und ihr Quelltext gespeichert. Aus dem Quelltext werden relevante Merkmale extrahiert, die Hinweise auf eine mögliche Suchmaschinenoptimierung geben. Zu diesen Informationen werden noch die Ladegeschwindigkeit der Seite und eine Kopie der robots.txt der Website erfasst.

2.2 Identifizierung von SEO-Plug-ins und Analytics Tools

In diesem Prozess wird der HTML-Code auf Hinweise für SEO-Plug-ins und Analytics Tools überprüft. Dafür werden manuell zusammengestellte Listen genutzt, die den Namen der Tools und ein Suchmuster enthalten.

2.3 Kategorisierung der URL

Die Kategorisierung der URL erfolgt anhand von vorab definierten Listen, die sich durch ihren Ergebnistyp und der Wahrscheinlichkeit von SEO auf diesen Typen voneinander abgrenzen. Diese Listen werden fortlaufend erweitert. Die Festlegung der Kategorien folgt der Annahme, dass insbesondere Seiten mit kommerziellen Absichten und große informationelle Angebote

wie Nachrichtendienste Suchmaschinenoptimierung einsetzen. Die Auswahl und Zusammenstellung der Seiten für die jeweiligen Kategorien basiert dabei auf den Top-Quellen in Suchergebnissen und auf der Identifizierung von SEO-Agenturkunden, die anhand von Kundenlisten dieser Agenturen erfolgte. Insgesamt erfolgte dieser Ansatz anhand der Annahme, dass die Diversität an Quellen in Top-Suchergebnissen relativ gering ist (Goel et al., 2010) und dazu die meisten Klicks nur auf eine geringe Anzahl von Domains erfolgen (Petrescu, 2014). Durch diese Vorgehensweise lässt sich die Vielzahl an Domains gut abdecken. Die Domains sind in folgende Kategorien eingeteilt und haben dabei den jeweils genannten Umfang an Seiten:

- nicht optimiert (1 Seite)
- Kunde von SEO-Agenturen (1.004 Seiten)
- Nachrichtendienst (1.203 Seiten)
- Websites mit Werbung (325 Seiten)
- Unternehmens-Webseite (72 Seiten)
- Online-Shops (178 Seiten)

2.4 Analyse der SEO-Indikatoren

Die ausgewählten Indikatoren für SEO basieren auf einer Sichtung der Fachliteratur (u. a. Enge, 2015; Erlhofer, 2019) und aus einer Befragung von SEO-Experten (Schultheiß/Lewandowski, 2020). Die Analyse erfolgt anhand folgender Kriterien, die auf ihr Vorkommen im Quelltext geprüft werden:

- Microdata-Formate
- Online-Werbung
- HTTPS
- SEO in robots.txt
- Sitemap
- Viewport
- Nofollow Tags
- Canonical Links
- Description Tags
- Title Tags.

2.5 Bewertung der SEO-Wahrscheinlichkeit

Für den regelbasierten Klassifikator sind folgende Kategorien und Regeln definiert:

- *Höchstwahrscheinlich optimiert:* Die Webseite ist höchstwahrscheinlich optimiert, wenn entweder ein SEO-Plug-in im Quelltext gefunden wurde, der Seitenbetreiber Kunde einer SEO-Agentur ist, die Seite ein Nachrichtenangebot ist, Werbeanzeigen auf der Seite sind oder mindestens ein Microdata-Format vorhanden ist.
- *Wahrscheinlich optimiert:* Die Seite ist wahrscheinlich optimiert, wenn sie nicht höchstwahrscheinlich optimiert ist und zumindest eines der folgenden Kriterien erfüllt: (1) Die Seite ist ein Online-Shop oder eine Unternehmensseite, (2) auf der Seite wurden Analytics-Tools identifiziert, (3) als Übertragungsprotokoll wird HTTPS eingesetzt, (4) SEO-spezifische Hinweise wurden in der robots.txt gefunden, (5) die Website hat eine Sitemap, (6) ein Viewport ist definiert, (7) es wurde ein Nofollow oder ein Canonical Link gefunden, (8) die Ladezeit liegt unter drei Sekunden.
- *Höchstwahrscheinlich nicht optimiert:* Die Domain der Seite ist auf der Liste mit nicht optimierten URLs.
- *Wahrscheinlich nicht optimiert:* Die Seite ist wahrscheinlich nicht optimiert, wenn sie nicht höchstwahrscheinlich nicht optimiert ist und wenn mindestens eines der folgenden Kriterien erfüllt ist: (1) kein Description Tag, (2) kein Title Tag, (3) wenn identische Title Tags auf Unterseiten sind und (4) keine Open Graph Tags definiert sind.

3 SEO-Tool

Das SEO-Tool ist eine Webanwendung, die die Durchführung der Prozesse demonstriert. Nach Eingabe einer URL wird die Analyse durchgeführt. Die Anwendung ist in Python entwickelt und nutzt Flask (in Python geschriebenes Web-Framework), Selenium (portables Framework zum Testen von Webanwendungen), Beautiful Soup (Paket zum Parsen von HTML und XML) und Pandas (Programmbibliothek zur Datenmanipulation und -analyse). Abbildung 2 zeigt ein Beispiel für einen generierten Ergebnisbericht. Der Bericht ist nach den Ergebnissen aus den Prozessen eingeteilt:

1. SEO-Bewertung: Wahrscheinlichkeit von SEO auf der gegebenen Webseite
2. Tools und Plug-ins: Verwendung von SEO-Tools und Analyse-Tools
3. URL Category: Ergebnis der Kategorisierung der URL
4. Indicators for SEO: Übersicht der Indikatoren für SEO.

Abb. 2 Ergebnisbericht im SEO-Tool

In den Spalten wird das Merkmal mit einer Checkbox und weiteren Erklärung angezeigt. Das Tool zeigt auch tiefer gehende Erklärungen zu allen Prozessen an, die durch Anklicken einer Kategorie geöffnet werden können. Schließlich kann der Bericht als CSV-Datei heruntergeladen werden.

4 Zusammenfassung und weiteres Vorgehen

Das SEO-Tool bietet die Möglichkeit einer Echtzeitanalyse zur Identifizierung von Suchmaschinenoptimierung auf einer Seite. Dabei basiert die Analyse bisher auf technischen Indikatoren und einer Kategorisierung der URL durch Abgleiche mit manuell zusammengestellten Listen, in denen wahrscheinlich optimierte und nicht optimierte Websites gespeichert sind. Diese Listen werden kontinuierlich aktualisiert. Die definierten halb-automatischen Prozesse werden ebenfalls stetig getestet und weiterentwickelt. Mithilfe des Tools können die entwickelten Prozesse und Regeln schnell überprüft werden. In dem weiteren Vorgehen werden die Indikatoren durch externe SEO-Signale wie die Anzahl der Backlinks auf einer Website erweitert. Zusätzlich werden Klassifikationen mit Machine-Learning-Algorithmen getestet, um zusätzliche Indikatoren zu ermitteln.

Forschungsdaten

Der Quelltext des Tools kann über die OSF-Plattform (dx.doi.org/10.17605/OSF.IO/ETZHD) abgerufen werden.

Förderung

Das Projekt „SEO-Effekt“, aus dem dieses Tool hervorgeht, wird von der Deutschen Forschungsgemeinschaft (DFG) unter der Projektnummer 417552432 gefördert.

Literaturverzeichnis

Enge, E.; Spencer, S.; Stricchiola, J. (2015): *The Art of SEO: Mastering Search Engine Optimization*. Sebastopol, CA: O’Reilly.

Erlhofer, S. (2019): *Suchmaschinen-Optimierung: Das umfassende Handbuch*. Bonn: Rheinwerk Verlag.

Goel, S.; Broder, A.; Gabrilovich, E.; Pang, B.(2010): Anatomy of the long tail. In: Davison, B. D.; Suel, T.; Craswell, N.; Liu, B. (Hrsg.): *Proceedings of the Third*

ACM International Conference on Web Search and Data Mining - WSDM '10. New York, NY: ACM Press.

Li, K.; Lin, M.; Lin, Z.; Xing, B. (2014): Running and Chasing – The Competition between Paid Search Marketing and Search Engine Optimization. In: *47th Hawaii International Conference on System Sciences, Waikoloa, HI*. IEEE, S. 3110–3119. https://doi.org/10.1109/HICSS.2014.640

McCue, T. (2018): SEO Industry Approaching $80 Billion But All You Want Is More Web Traffic. *Forbes*. https://www.forbes.com/sites/tjmccue/2018/07/30/seo-industry-approaching-80-billion-but-all-you-want-is-more-web-traffic/

Petrescu, P. (2014): Google Organic Click-Through Rates in 2014. https://moz.com/blog/google-organic-click-through-rates-in-2014

Schultheiß, S.; Lewandowski, D. (2020): "Outside the industry, nobody knows what we do". SEO as seen by search engine optimizers and content providers. *Journal of Documentation* 77 (2), 542–557. https://doi.org/10.1108/JD-07-2020-0127

In: T. Schmidt, C. Wolff (Eds.): Information between Data and Knowledge. Information Science and its Neighbors from Data Science to Digital Humanities. Proceedings of the 16th International Symposium of Information Science (ISI 2021), Regensburg, Germany, 8th–10th March 2021. Glückstadt: Verlag Werner Hülsbusch, pp. 299–306. DOI: doi.org/10.5283/epub.44949.